Methods for the Prediction of Complex Biomolecular Structures

Inaugural-Dissertation
zur
Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für
Biologie
an der

Universität Duisburg-Essen

vorgelegt von

Christoph Wilms

aus Essen

Oktober 2013

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatik der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann

2. Gutachter: Prof. Dr. Peter Bayer

Vorsitzende des Prüfungsausschusses: Prof. Dr. Angela Sandmann

Tag der mündlichen Prüfung: 10. Februar 2014

Zusammenfassung

Die erste hochaufgelöste Struktur eines Proteins wurde 1985 von John Kendrew und Max Perutz aufgelöst. Seitdem ist die experimentelle Aufklärung ein wichtiger Bestandteil der biologischen Forschung. Allerdings ist die Aufklärung der Strukturen von biomolekularen Komplexen sehr schwierig. Diese Strukturen sind jedoch immens wichtig für das Verständnis vieler biologischer Phänomene auf molekularer Ebene. Aus diesem Grund hat sich ein Forschungsfeld entwickelt, das computergestützte Modellierung zur Vorhersage von biomolekularen Strukturen verwendet.

In dieser Promotionsschrift sollten Methoden zur Vorhersage von komplexen biomolekularen Strukturen entwickelt werden. Diese Methoden basieren auf drei unterschiedlichen Ansätzen:

Die erste Methode wurde für Proteine entwickelt, die aus mehreren Domänen bestehen. Die Methode nutzt vorhandene Strukturen der einzelnen Domänen und experimentelle Daten, die geometrische Relationen der Domänen abbilden, und ermöglicht die Untersuchung konformationeller Änderungen bedingt durch äußere Einflüsse, wie beispielsweise das Zuführen eines Substrates. Als Fallbeispiel wurde die Konformation des flexiblen zwei-Domänen Proteins peptidylprolyl cis/trans isomerase NIMA-interacting 1 (Pin1) untersucht, sowie die Änderung als Reaktion auf die Zugabe des Substrates polyethylene glycol (PEG).

Die zweite Methode basiert auf dem neuen Verfahren Direct Coupling Analysis (DCA), das es ermöglicht geometrische Kontakte von Aminosäuren anhand eines multiplen Sequenzalignments (MSA) vorherzusagen. DCA nutzt eine Korrektur zur Vermeidung einer Stichprobenverzerrung bedingt durch die Auswahl der Sequenzen für das MSA. Die hier vorgestellte Optimierung ermöglicht eine robustere Vorhersage der geometrischen Kontakte. Die optimierte Methode wurde für die Analyse von Human Immunodeficiency Virus-1 Envelope Protein (HIV-1 Env) eingesetzt.

Die letzte Methode wurde entwickelt, um Binderegionen des negativ geladenen Heparansulfates an Proteinen vorherzusagen. Dafür haben wir ein Modell entwickelt, das auf der elektrostatischen Wechselwirkung basiert. Die Fallbeispiele sind hier verschiedene Heparansulfat bindenden Proteine, wie das Chemokine CCL3 und den Hedgehog Proteinen.
Insgesamt wird gezeigt, dass für verschiedene Arten von biomolekularer Strukturen und Komplexe moderne computergestützte Methoden Einsichten liefern, die im Einklang mit Experimenten stehen.

Contents

List of Abbreviations

CCR5 C-C Chemokine Receptor Type 5

CD4 Cluster of Differentiation 4 Receptor

CXCR4 C-X-C Chemokine Receptor Type 4

DCA Direct Coupling Analysis

DFT Discrete Fourier Transformation

Dhh Desert Hedgehog

DI Direct Information

FFT Fast Fourier Transformation

GAG Glycosaminoglycan

GP41 Glycoprotein 41

GP120 Glycoprotein 120

GP160 Glycoprotein 160

Hh Hedgehog

HIV Human Immunodeficiency Virus

HIV-1 Env Human Immunodeficiency Virus-1 Envelope Protein

HMM Hidden Markov Model

HS heparan sulfate

HSPG heparan sulfate proteoglycan

HSQC Hetero Single Quantum Coherence

Ihh Indian Hedgehog

MD Molecular Dynamics

MI Mutual Information

MSA Multiple Sequence Alignment

NMR Nuclear Magnetic Resonance

NOE Nuclear Overhauser Effect

PDB Protein Data Bank

PEG polyethylene glycol

Pfam Protein families

Pin1 peptidylprolyl cis/trans isomerase NIMA-interacting 1

PPIase peptidylprolyl isomerase

PRE Paramagnetic Relaxation Enhancement

Ptc Patched

Smo Smoothend

Shh Sonic Hedgehog

TP true positive

Ttv Tout Velu

List of Figures

List of Tables

1

Introduction

Since the first protein structure was solved by Kendrew et al. [59] and Perutz et al. [107] the experimental techniques have been refined, and more than 86 000 structures are available in the Protein Data Bank (PDB) today. Only the structure of a protein enables scientists to fully understand its properties and functions. Thus, the demand for new structures is ever growing. Today, X-ray and Nuclear Magnetic Resonance (NMR) experiments are the major source of new structures. However, these experiments are very demanding regarding the time to produce the purified protein crystal (X-ray) or purified protein solution (NMR) and therefore very cost-intensive. A field of research has evolved around the idea to use computer simulations and algorithms for the prediction of protein structures and complexes [16, 25, 57, 72, 96, 115, 152].

Although progress has been achieved, the current state-of-the-art methods still have difficulties regarding the handling of flexible structures and structures with non-protein components. Further methods are required to handle the massively growing amount of sequence data and other experimental data that could potentially be translated into structural information.

Therefore, we developed three methods that are independent of each other and have no thematic overlap. The methods are described in separate chapters. Each chapter starts with an introduction outlining the current state of research. Next, the methods are explained in detail, followed by the results and discussion of the case studies. Finally, each chapter contains a summary combined with an outlook.

The first case study is an example for the refinement of a flexible two-domain protein using experimental data. We analyzed the conformation of native peptidylprolyl cis/trans

isomerase NIMA-interacting 1 (Pin1) and its conformational response upon the addition of polyethylene glycol (PEG).

The second chapter is about optimization for the recently published Direct Coupling Analysis (DCA) method [95]. The DCA method has been developed for the inference of direct contacts from a Multiple Sequence Alignment (MSA) and is able to use the massively growing amount of sequence data for the refinement of protein structures. We applied this method for the analysis of Human Immunodeficiency Virus-1 Envelope Protein (HIV-1 Env).

In the last chapter, we present a method for the prediction of heparan sulfate interacting regions of proteins. This method refers to the analysis of complexes with non-protein components.

2

Using Paramagnetic Relaxation Enhancement data for the structural characterization of the flexible two-domain protein Pin1

This chapter deals with a new method to address the problem of characterizing flexible structures, and we applied this method to Pin1, an evolutionary conserved two-domain protein.

2.1 Pin1

The enzyme peptidylprolyl cis/trans isomerase NIMA-interacting 1 (Pin1) consists of a C-terminal peptidylprolyl isomerase (PPIase) and an N-terminal WW domain [84]. The PPIase catalyzes the cis-trans isomerization of prolines [31], which have two distinct conformational states of the backbone peptide bond. These two states are separated by a large energy barrier and thus the interconversion is a slow process [32, 137]. WW domains derive their name from the two highly conserved tryptophans, which bind proline-rich or phosphoserine- phosphothreonine-containing motifs [15, 113]. Hence, Pin1 isomerizes specific phosphorylated serine/threonine-proline bonds [77] and therefore regulates the function of phosphoproteins [148] increasing the rate of cis-trans interconversion by a

factor of 1000 [104]. The catalytic activity resides solely in the PPIase domain and the WW domain has no catalytic activity at all.

The cis-trans isomerization of prolines can be considered as a time-limiting step in folding [8, 121, 122]. Thus, Pin1 can be interpreted as a molecular switch, which determines the pathways of many phosphoproteins [77]. Furthermore, it was also suggested that Pin1 acts as a molecular timer because it interacts with many cell cycle regulatory proteins [150]. The multiplicity of interaction partners makes Pin1 an interesting protein as it has been related to cancer and Alzheimer's and Parkinson's disease [150].

The fold of the PPIase-domain is typical for parvulin-like proteins. Its core consists of a four-stranded antiparallel β-sheet surrounded by four α-helices. The WW domain consists of a triple stranded β-sheet. Several structures of Pin1 have been deposited in the PDB [6]. In general, structures derived from x-ray experiments (e.g. 1pin) show tight packing of the two domains [113, 134, 153], whereas structures solved by NMR (e.g. 1nmv) show a high flexibility of the WW domain for which no inter-domain Nuclear Overhauser Effects (NOEs) could be detected [3, 52]. Interestingly, various crystal structures contain a PEG moiety located at the domain interface (e.g. 1pin, 3tdb). Figures 2.1b and 2.1a illustrate the different conformations. It has been reported that the two domains behave independently but can also act as a single entity upon substrate binding [52]. In this study, we analyzed the conformational response of Pin1 upon the addition of PEG.

(a) 1pin **(b)** 1nmv (NMR model 1)

Figure 2.1: Open and closed conformation of Pin1 - Comparison of two conformational states of Pin1. In x-ray structure 1pin the WW domain resides close to the PPIase domain, whereas in the NMR structure 1nmv the WW domain is located away from the PPIase-domain. PDB structure 1pin contains a chain break between residues $40 - 44$, because these residues could not be determined from the electron density maps [113].

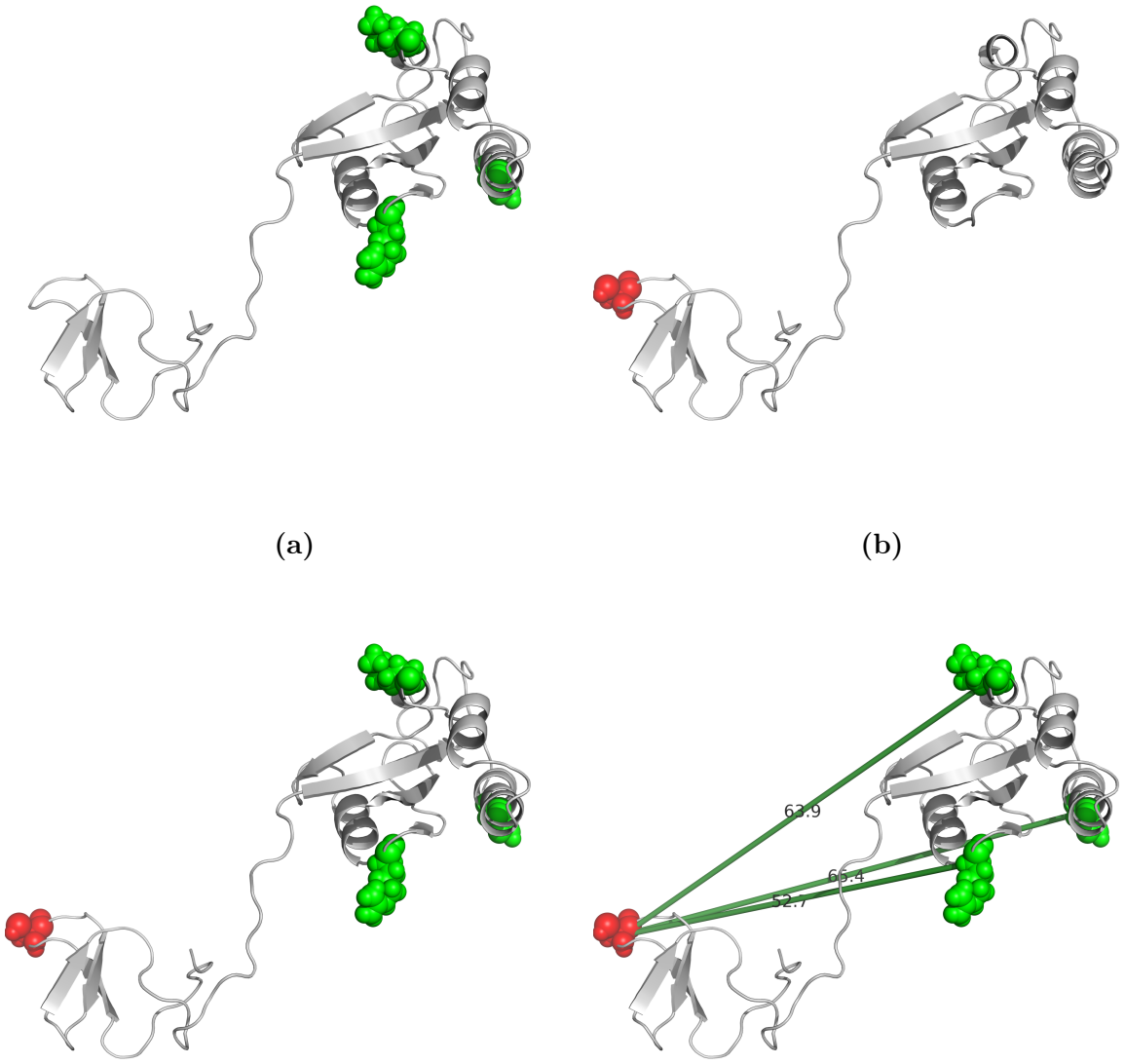
2.2 Methods

In this section, we present the methods on which the Pareto optimization is based. At first, the experiment, which measured the geometrical relations among the two domains, is outlined. Afterwards, we describe the Pareto optimization in detail.

2.2.1 Paramagnetic Relaxation Enhancement

For the analysis of Pin1, a Paramagnetic Relaxation Enhancement (PRE) experiment was performed by Peter Bayer, Research Group Structural and Medicinal Biochemistry, University of Duisburg-Essen, to analyze the effect of PEG on the domain interaction.
Starting from human Pin1 the two endogenous cysteine residues C57 and C113 were replaced with alanine through site-directed mutagenesis and a new cysteine replaced S18. Next, a paramagnetic pyrrolidinyloxy (PROXYL) moiety label was covalently attached to C18 via covalent coupling.
PRE is a Nuclear Magnetic Resonance (NMR) method. NMR is based on the Zeeman splitting of the energy levels of NMR-active nuclei in the presence of a magnetic field. A torque acts on the magnetic moments of the nuclei and induces a precession around the axis of the magnetic field with a frequency that is proportional to the strength of the magnetic field. A paramagnetic substance induces local distortions in the magnetic field. The closer a NMR-active nuclei is to the paramagnetic substance, the stronger the local change of the magnetic field and thus the change in its frequency. The measured signal intensities are very sensitive to a change in the frequencies and any change results in a lower signal. Therefore, a decrease of the signal intensities can be related to the relative distance of the NMR-active nuclei and the paramagnetic substance.
The transformation of PRE data to distance constraints is illustrated schematically in figure 2.2. At first, peak intensities I_{red} are calculated from the ^{1}H $-^{15}$ $N-$Hetero Single Quantum Coherence (HSQC) spectra for the unlabeled Pin1 variant (figure 2.2a). After labeling Pin1 with the PROXYL-iodoacetamide (figure 2.2b), peak intensities I_{ox} are calculated from the HSQC spectra for the labelled variant (figure 2.2c). PREs are then calculated from the signal height ratio I_{ox}/I_{red} of the extracted peak intensities. In principle, one could transform the PREs into distance constraints and use these constraints to create a structural model [2, 85] (figure 2.2d). Due to the very high motional flexibility

of the system this was not possible. Therefore, we employed Pareto optimization to infer structural information from the experimental data. The method is described in the next subsection. The experiment was repeated after adding PEG with a molecular weight of 400 g/mol. From both experiments, with and without PEG, we could infer positions of the labelled residue, which best explained the experimental data. From the geometrical relation between the results with and without PEG we can infer an influence of PEG upon the system.



(a) **(b)**

(c) **(d)**

Figure 2.2: Transformation of PRE data to distance constraints - At first, the peak intensities I_{red} are measured for residues of the wild type (shown as green spheres in (a)). In the next step, the wild type is labelled with the paramagnetic substance (shown as red spheres in (b)), and the peak intensities I_{ox} are measured for the same residues (c). In principle, the height of the signal ratio I_{ox}/I_{red} can be transformed to distance constraints (shown as green lines in (d)).

2.2.2 Pareto optimization of the experimental data

Physical properties of the paramagnetic label and the experimental data were used as the basis for two heuristic optimization functions generating constraints for possible positions of the paramagnetic label around the PPIase. The dependence of the signal ratio I_{ox}/I_{red} upon the distance to the paramagnetic label is schematically shown in figure 2.3. The closer the NMR-active nuclei is to the paramagnetic label, the stronger its signal is decreased until it is too weak to be detected. The border between the region where no signal is received and the NMR-active region is defined as the extinction radius.

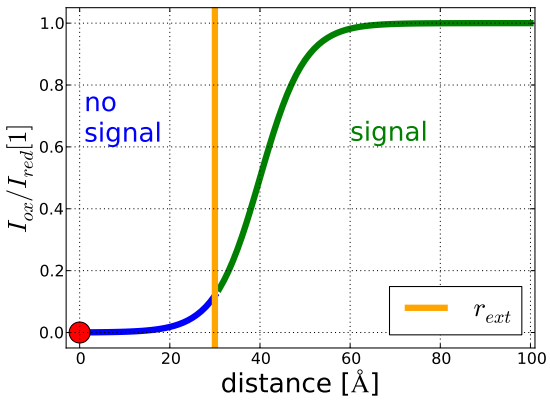


Figure 2.3: Definition of the extinction radius - Below a certain threshold (illustrated as an orange line) the influence of the paramagnetic label (shown as a red dot) is too strong, and no signal can be detected ($I_{ox} \approx 0$, shown as a blue line). For distances above the threshold, the signal can be detected ($I_{ox} \approx I_{red}$, shown as a green line). The threshold defines the extinction radius r_{ext}.

The first heuristic optimization function f_1 is based on the extinction radius. The function counts the number of residues for which signals were experimentally observed (green) that lie within the extinction radius around the paramagnetic label (red) in the sampled conformation, and thus should not give signals, and the number of residues for which no signals were experimentally observed (blue) lying outside the extinction radius. The optimization function is illustrated in figure 2.4a for a conformation in conflict with the experiment. A conformation in agreement with the experiment minimizes this count and is shown in figure 2.4b.

The second optimization function is based on the physical principle that the recovery of the signal monotonously increases with distance to the paramagnetic label. For each conformation, we calculated Spearman's rank correlation ρ between the distance from the paramagnetic label in the sampled conformation and the experimentally observed degree of extinction. A conformation in conflict with the experimental results yields a low ρ as illustrated in figure 2.4c, whereas a conformation in agreement with the experimental results maximizes Spearman's rank correlation and yields $\rho = 1$ as illustrated in figure 2.4d. We defined the second optimization function as $f_2 = 1 - \rho$ and thus for the second function we also have a minimization problem.

For the identification of optimal arrangements of PPIase and WW domain that best explained experimental data we used Pareto dominance as illustrated in figure 2.5. In multi-objective problems, a solution can be treated as a vector $\vec{a} = (a_1, ..., a_N)$ consisting of the results a_i from the different optimization functions, which constitute the search space $\mathcal{S}$. A vector $\vec{b}$ dominates vector $\vec{a}$, mathematically expressed by $\vec{b} \prec \vec{a}$ [20], if

$$\forall i \in \{1, ..., m\} : b_i \leq a_i \text{ and } \exists\, j \in \{1, ..., m\} : b_j < a_j. \tag{2.1}$$

Pareto dominance is a technique applied to multi-objective problems to identify non-dominated solutions, which are referred to as Pareto optimal. The set $\mathcal{S}$ of Pareto optimal solutions is defined as

$$\{\vec{x} \in \mathcal{S} \mid \nexists\, \vec{y} \in \mathcal{S} : \vec{y} \prec \vec{x}\}. \tag{2.2}$$

Optimal relative arrangements of PPIase and WW domains were determined in two steps. First promising regions of the paramagnetic label attached to the WW domain around the PPIase were located using a coarse grid with spacing of 0.1 nm and 100 grid points along each of the three Cartesian coordinate axes centered on the PPIase. Promising regions were refined in the second step using a grid of 0.05 nm spacing and again 100 grid points in each of the three directions.

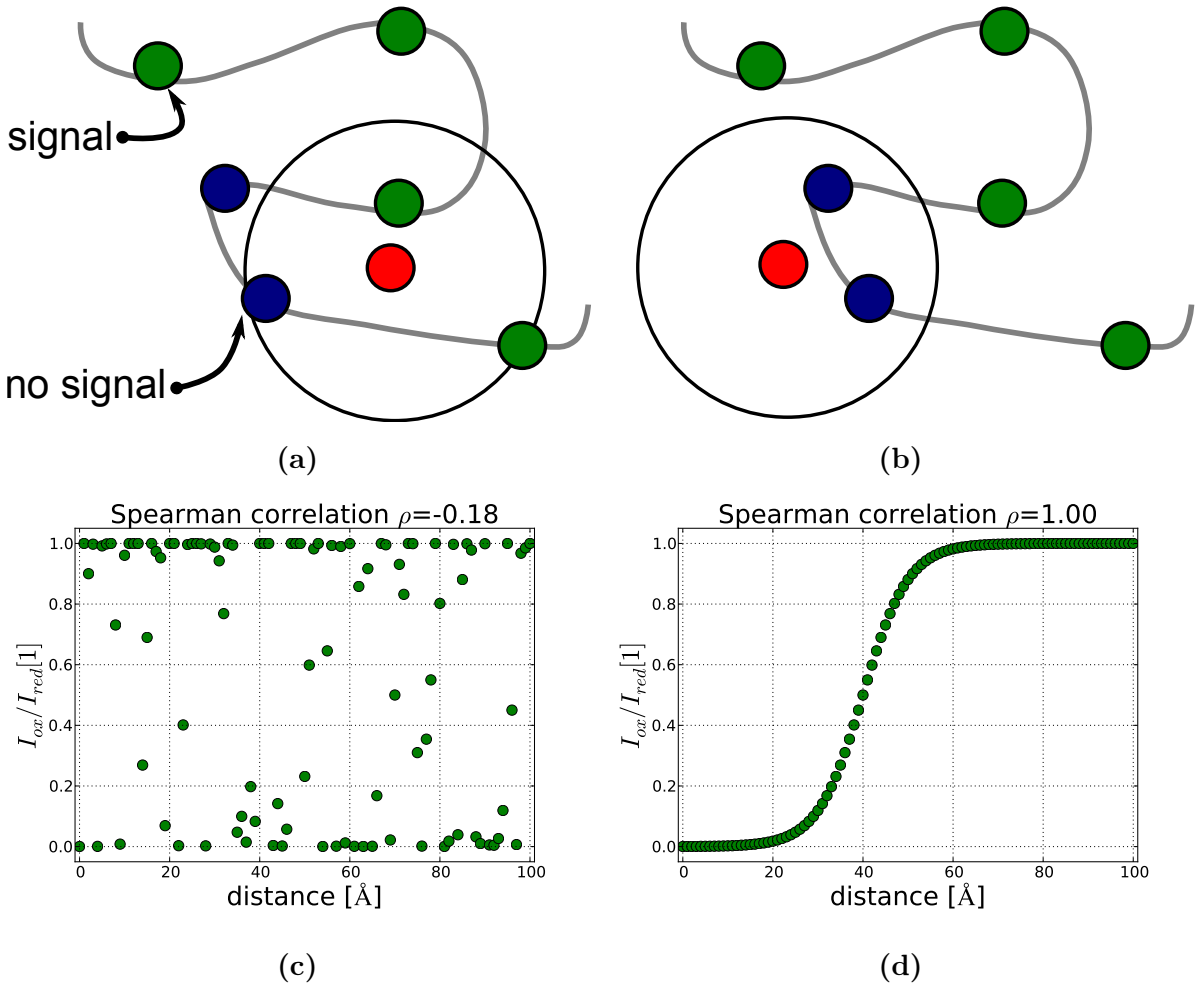


Figure 2.4: Optimization functions used for the Pareto optimization - The first optimization function is shown in figures (a) and (b). Figure (a) illustrates a conformation in conflict with the experimental data ($f_1 = 3$), whereas figure (b) shows a conformation in perfect agreement ($f_1 = 0$). The second function is illustrated in figure (c) for a conformation in conflict (low ρ) and in figure (d) for a conformation in perfect agreement ($\rho = 1$) with the experimental data.

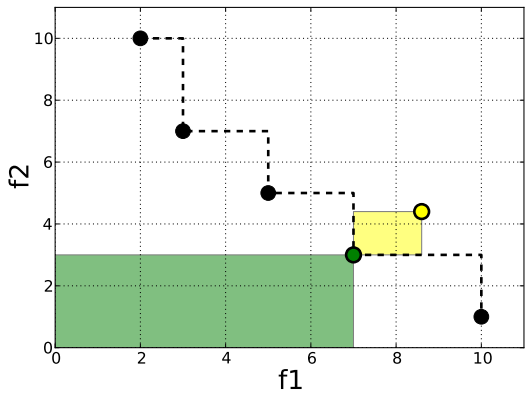


Figure 2.5: Pareto principle - A solution is Pareto-optimal if no other solution dominates any of its optimization functions. The yellow dot is not Pareto-optimal because it is dominated by the green solution, which lies inside the area of the yellow rectangle meaning that any of its optimization functions are lower than any of the yellow's. The Pareto front consists of the green and black dots.

2.3 Results

First, we used the WW domain of PDB structure 1pin [113] for the determination of the extinction radius. The extinction radius is illustrated in figure 2.6 on the WW domain. We assumed that the conformation of the WW domain does not change much whether it is close to the PPIase or not. The distribution of the non-signal- and signal-giving residues in figure 2.6 shows that the chosen conformation is in good agreement with the experimental data. The non-signal-giving residues (blue) are close to the paramagnetic substance (red), and the signal-giving residues (green) are on the opposite site. From this data, we determined the extinction radii around the paramagnetic label as 2.09 nm in the presence of PEG, and as 2.25 nm in the absence of PEG. The experimentally derived extinction radii are in agreement with the range of 2 to 2.5 nm reported in the literature [38].

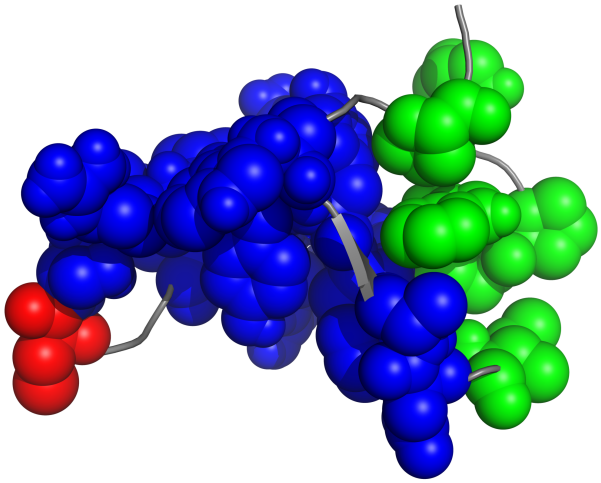


Figure 2.6: Using the WW domain to determine r_{ext} - The paramagnetic label is attached to C18, which is shown as red spheres. Residues for which no signal could be detected are shown as blue spheres and residues with a signal are shown as green spheres.

Pareto Optimization of the two optimization functions yields positions for the paramagnetic label and thus also for the cysteine to which it is attached. The results of

the two optimization functions for each conformation in the presence of PEG are shown in figure 2.7, with the Pareto-optimal points highlighted as red bullets.

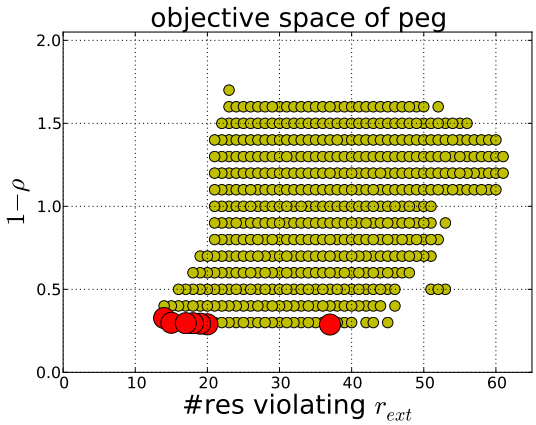


Figure 2.7: Pareto dominance - Illustration of the values of the two optimization functions for each conformation based on the refinement in the presence of PEG. The Pareto-optimal solutions are shown as red dots.

The solutions for the positions of the paramagnetic label around Pin1 are shown in figure 2.8. The analysis showed that the addition of PEG decreased the distances between the PPIase, and the WW domain and tightened the domain interaction.
We provided further results supporting our Pareto results in our publication [90]. We observed a temperature-dependent splitting of resonances for residues located in the second loop of the WW domain, which we traced back to the existence of two states, which we refer to as A and B. Without PEG, the system favors the more open A state. However, by adding PEG the population of the two states is reversed and the system favors the more closed B state.
We used Molecular Dynamics (MD) simulations of full length Pin1 with and without a PEG molecule, and showed that the distance between S18 and various residues within the PPIase domain were shorter in the presence of PEG than in its absence. We could also show that the ligand-binding properties of Pin1 were affected by the presence of PEG. The open conformation A binds with less affinity to the peptide substrate than the closed conformation B. By modulating the binding affinity, the small molecule PEG can be used to control the catalytic activity of a whole protein.

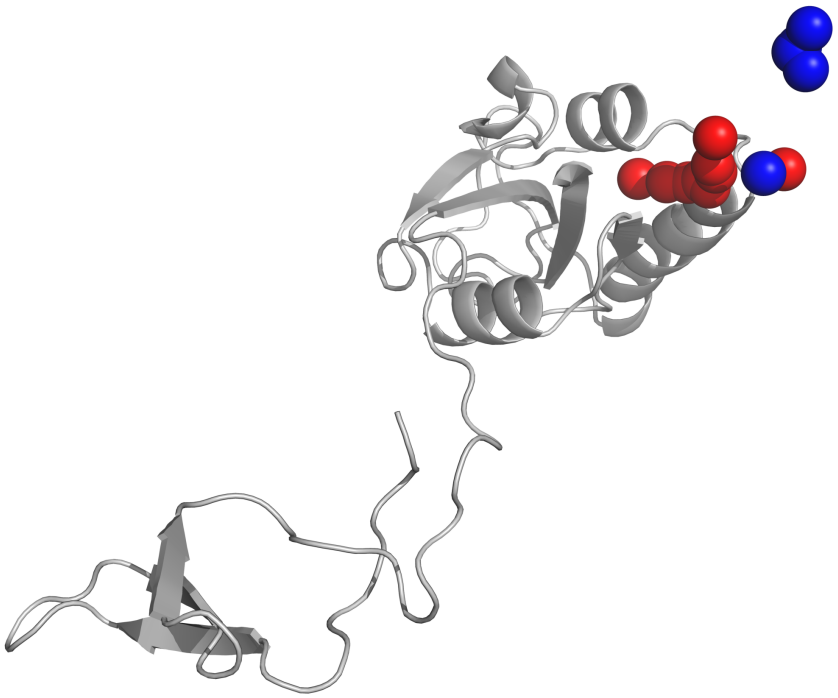


Figure 2.8: Pareto solutions around Pin1 - Pareto solutions in the absence (blue) and presence (red) of PEG.

2.4 Summary and outlook

The algorithm proved useful in the analysis of the PRE experiment. We identified two different distance regions of the WW domain relative to the PPIase, which we related to two different states A and B, with an open conformation A and a more closed conformation B. The addition of PEG decreased the distance between the two domains and thus favored state B. We also found that the two states have different binding affinities to the peptide substrate. Therefore, PEG can be used to modulate the catalytic activity of Pin1.
The accuracy of the algorithm is, of course, not comparable to the results one might have obtained by transforming the intensity ratios into distance constraints. However, sometimes the transformation of experimental data to distance constraint is not possible as was the case in the present study. In this case, the algorithm can be very useful. Because of the generic nature of the Pareto principle, the algorithm can be easily extended to various other problems. The only requirement is that one can formulate heuristic optimization functions that rely on experimental data and/or physical properties.

3

Optimizing the re-weighting method of Direct Information computation

This chapter deals with optimization for the re-weighting method applied in Direct Information (DI) computation. Direct Information is able to use the massively growing amount of sequence data for the characterization of protein structures. We also applied this method to the Human Immunodeficiency Virus-1 Envelope Protein.

3.1 Direct Information

Recently, the concept of Direct Information (DI) was introduced by Weigt et al. [139] and significantly improved by Morcos et al. [95]. The latter referred to the method as Direct Coupling Analysis (DCA). Here we refer to the method as DI as it is the quantity of interest. DI uses the evolutionary information contained in a Multiple Sequence Alignment (MSA) of a given protein family. An example of an MSA is illustrated in figure 3.1.

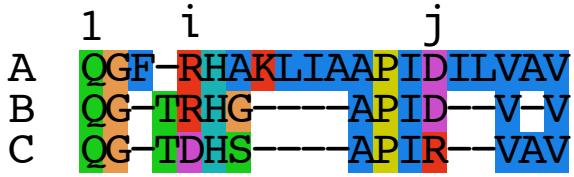


Figure 3.1: MSA example - MSA of three sequences A, B and C. Coevolution can be observed in columns i and j.

The amino acid composition of a protein may change during evolution, although its fold remains almost unchanged. The occurring changes in the sequences of the MSA can be used to infer structural or functional information of the involved proteins. For example, a conserved residue can be considered important for the structure or function of the protein. Although this information in itself is already useful for the understanding of the protein, the contained information is sparse regarding the actual structure of the protein. To get a better understanding of the protein, one would like to infer geometrical relations between pairs of residues. One way to accomplish this is by searching for correlated evolution, also known as coevolution [23]. Imagine two residues distant in sequence, but close in tertiary structure. The first residue is the positively charged arginine (R) and the second residue is the negatively charged aspartic acid (D) forming a salt bridge. If the first residue would also be mutated to aspartic acid the salt bridge cannot be formed, and the stability of the protein is decreased. However, if at the same time the second residue mutates to arginine, this compensatory mutation would probably be as stable as the wild type. This coevolution can be seen in the example MSA in figure 3.1. The coevolution takes place in columns i and j.

Coevolution often takes place between residues which are close in space [33, 108, 149]. Hence, the identification of those coevolving residues is of great interest. The coevolving residue pair can be transformed to a structural constraint, which then can be used to infer properties of the protein fold.

Several studies tried to develop methods for the inference of interacting residues from MSAs, e.g.[9, 39, 56, 64, 70, 82, 89, 98, 130]. One of the methods is called Mutual Information (MI), which measures the mutual dependence of two columns of the MSA. The mathematical details are given in subsection 3.2.1. The main problem of MI is its inability to distinguish direct from indirect couplings. For example, if residue i is coupled with residue j, but j is also coupled with residue k, then there will be also a high MI for the pair (i, k) (compare figure 3.2). The disentangling of direct from indirect couplings is the huge success of DI. The theory of DI is explained in subsection 3.2.2. The power of DI has already been shown in an extensive study by Morcos et al. [95].

The quality of these methods depends strongly on the quality of the underlying MSAs. All methods based on MSAs require them to be “well sampled”, which means that there should be a large number of sequences with a low sequence identity among the sequences.

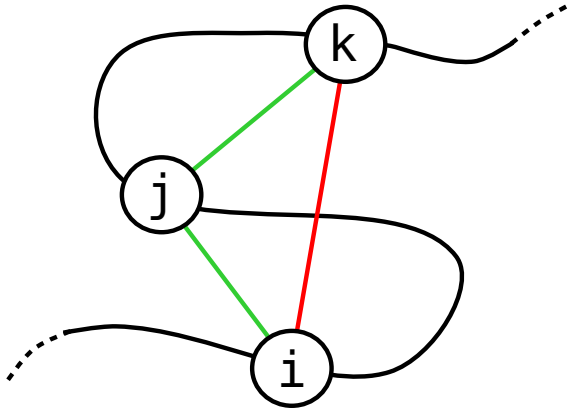


Figure 3.2: Direct and indirect couplings - Residue pairs (i, j) and (j, k) are in direct contact, indicated by green lines. The red line illustrates the indirect coupling of residues i and k through residue j.

For an alphabet size $q = 21$ at least $q^2 = 441$ sequences are required to be in principle able to sample all possible combinations of amino acids.
Recent technological advances in the sequencing field [81] have boosted the number of available sequences. Any selection of sequences is in general prone to a sampling bias. Therefore, Procaccini et al. [110] introduced a re-weighting strategy for the correction of sampling biases. In this study we analyzed the influence of re-weighting upon DI applied to the dataset used by Morcos et al. [95].
We selected the Human Immunodeficiency Virus-1 Envelope Protein as an example for the application of DI as a tool to analyze the structure and function of a protein. One might ask why we did not apply the method to other proteins featured in this work, namely Pin1, CCL3 and Hedgehog. Not enough sequences are available for Pin1 and Hedgehog (< 170 sequences for Pin1 and < 250 for Hedgehog). In the case of CCL3 and Hedgehog we are interested in the interaction with heparan sulfate, and DI cannot characterize the interaction with heparan sulfate as it is not part of the MSA. In principle, one could look for indirect effects caused by the interaction with heparan sulfate. Because CCL3 is a rather short protein with a sequence length ≈ 60 amino acids, we were not able to extract couplings related to the interaction with heparan sulfate.

3.2 Methods

In this section, the underlying theory of DI is presented. However, first the theory of MI is outlined as it is also part of the DI algorithm.

3.2.1 On the theory of Mutual Information

MI works on sequences aligned in an MSA, with M sequences and N positions in each sequence. The alignment contains the 20 standard amino acids and a character representing gaps, resulting in an alphabet of $q = 21$ different characters, which are translated to consecutive numbers $1, \ldots, q$.
In the first step single site and pair frequencies are calculated,

$$f_i(A) = \frac{1}{M} \sum_{a=1}^{M} \delta_{A,A_i^a} \tag{3.1}$$

$$f_{ij}(A,B) = \frac{1}{M} \sum_{a=1}^{M} \delta_{A,A_i^a} \delta_{B,A_j^a}, \tag{3.2}$$

with $1 \leq i, j \leq L$ iterating through the alignment positions, $1 \leq A, B \leq q$ enumerating the alphabet and δ representing the Kronecker symbol, which equals one if the indices are the same and equals zero otherwise. The single site frequency $f_i(A)$ counts the occurrences of amino acid A in the column i and the pair frequency $f_{ij}(A,B)$ counts all simultaneous occurrences of amino acid A at position i and B at position j. The observed frequencies can be interpreted as probabilities. The uncertainty of the single site and pair probabilities can be estimated using Shanon entropy $H(i)$ and the joint entropy $H(i,j)$

$$H(i) = -\sum_{A=1}^{q} p_i(A) \log_q p_i(A) \tag{3.3}$$

$$H(i,j) = -\sum_{A=1}^{q} \sum_{B=1}^{q} p_{ij}(A,B) \log_q p_{ij}(A,B). \tag{3.4}$$

The Mutual Information of columns i and j is defined as

$$MI(i,j) = H(i) + H(j) - H(i,j). \tag{3.5}$$

which can be transformed to

$$MI(i,j) = \sum_{A=1}^{q} \sum_{B=1}^{q} p_{ij}(A,B) \log_q \frac{p_{ij}(A,B)}{p_i(A)p_j(B)}. \tag{3.6}$$

MI ranges between 0 and 1. Two independent columns A and B have an MI equal to zero if column A does not contain any information about column B. If column A and B are dependent, e.g. they are identical or contain coevolving amino acids, their MI is close to one.

3.2.2 On the theory of Direct Information

This work focused on the optimization of the re-weighting algorithm used for DI computation. Therefore, the full mathematical description of DI is not part of this work. The mathematical description has been published by Morcos et al. [95] in their supplementary. Here, we limited the description to the most relevant formulas.
Just like MI, DI also works on sequences aligned in an MSA, with M sequences and N positions in each sequence. Again the alphabet contains the 20 standard amino acids and a character representing gaps, which are translated to consecutive numbers $1, \ldots, q$. Single site and pair frequencies are calculated according to equations 3.1 and 3.2. Due to phylogenetic relations among the sequences and a possible bias in the selection of the sequences, these counts may inherit a systematic error from the MSA. This error can be corrected by re-weighting the frequencies [110]. For each sequence A^a the number of similar sequences m^a is calculated as

$$m^a = \sum_{b=1}^{M} c_b, \quad c_b = \begin{cases} 1, & \iota(A^a, A^b) \geq xN \\ 0 \end{cases} . \tag{3.7}$$

with the fraction of identical sequence positions ι of two sequences and the identity threshold $0 < x < 1$. The frequencies are weighted with the factor $1/m^a$, which equals one for counts of sequences without similar sequences and down-weights counts of sequences for which similar sequences exist. The underlying assumption is that sequences with a low sequence identity carry more independent information. Low values of x down-weight densely sampled regions, whereas less densely sampled regions get a higher weight. All weights add up to an effective number of independent sequences $M_{eff} = \sum_{a=1}^{M} 1/m^a$. The re-weighted frequencies are given as

$$f_i(A) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \right) \tag{3.8}$$

$$f_{ij}(A,B) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \delta_{B,A_j^a} \right), \tag{3.9}$$

with a pseudo count $\lambda = wM_{eff}$ [26], where w is the weight of the pseudo count, for ensuring $f_{ij}(A,B) > 0 \;\; \forall \; A, B$ to avoid divergent couplings for frequency counts equal to zero.

The frequency counts can be used for the definition of a connected-correlation matrix C_{ij}

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B). \tag{3.10}$$

The off-diagonal elements of the inverse $(C^{-1})_{ij}$ contain the couplings $e_{ij}(A,B)$

$$e_{ij}(A,B) = -\left(C^{-1}\right)_{ij}(A,B). \tag{3.11}$$

From these couplings, an isolated two-site model $P_{ij}^{dir}(A,B)$ can be defined as

$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z}\exp\left(e_{ij}(A,B) + h_i(A) + h_j(B)\right), \tag{3.12}$$

with the auxiliary fields h and the partition function Z. The auxiliary fields are given by the three constraints

$$f_i(A) = \sum_{B=1}^{q} P_{ij}^{(dir)}(A,B) \tag{3.13}$$

$$f_j(B) = \sum_{A=1}^{q} P_{ij}^{(dir)}(A,B) \tag{3.14}$$

$$h_i(q) = h_j(q) = 0. \tag{3.15}$$

The partition function is defined as

$$Z = \sum_{A_i|i=1,\ldots,L} \exp\left(\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right). \tag{3.16}$$

The DI of alignment columns i and j is defined as the MI based on the isolated two-site model $P_{ij}^{(dir)}$ instead of the empirical pair frequencies

$$DI(i,j) = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B)\log_q \frac{P_{ij}^{(dir)}(A,B)}{f_i(A)f_j(B)}. \tag{3.17}$$

3.3 Improving the re-weighting algorithm

In this section the dependence of the re-weighting upon the threshold x in equation 3.7 is analyzed. Procaccini et al. [110] and Morcos et al. [95] stated that re-weighting is always superior to no re-weighting and varying the threshold from 70 % to 90 % has no significant effect.

(a) ι_{MSA} **(b)** ι_{PW}

Figure 3.3: Re-weighting example ι_{MSA} vs ι_{PW} - In the case of ι_{MSA} all gaps are used for the calculation of the pairwise sequence identities, whereas in the case of ι_{PW} gaps occurring in both sequences are removed. Here: $\iota_{MSA}(B, C) = 0.8$ and $\iota_{PW}(B', C') = 0.69$.

The re-weighting scheme by Procaccini et al. [110] and Morcos et al. [95] calculates the sequence identity based on rows of the MSA (ι_{MSA} in equation 3.7), which can lead to the inclusion of many gaps and thus sequences are more similar. We introduced a different re-weighting scheme ι_{PW}, which neglects gaps occurring in both sequences at the same position. This can be considered as a pairwise alignment of the two sequences and allows for a finer discrimination between similar and non-similar sequences. According to this definition it is always $\iota_{MSA} \geq \iota_{PW}$. Figure 3.3 illustrates the difference of the two re-weightings for sequences B and C of the MSA example in figure 3.1. Due to the insertions in sequence A, sequences B and C contain several gaps. These gaps increase the identity of the two sequences ($\iota_{MSA} = 0.8 > \iota_{PW} = 0.69$).
In the following DI_{MSA} (MI_{MSA}) refers to Direct Information (Mutual Information) based on the re-weighting applied by Morcos et al. [95] using rows of the MSA. DI_{PW} (MI_{PW}) refers to the Direct Information (Mutual Information) based on the re-weighting strategy, which uses "pairwise alignments" for the calculation of the sequence identities ι_{PW}. We compared the performance of both re-weightings using the dataset provided by Morcos et al. [95].

3.3.1 Implementation of the algorithm

The original code of DCA has been implemented in MATLAB [91], which we kindly received from Martin Weigt, but it is also available online[1]. We implemented the algorithm in Python[2] using Cython [4] to generate C-Extensions to speed up the calculations and NumPy [100] for N-dimensional array objects. The Python code is part of our "epitopsy"-libraray[3].

3.3.2 Data extraction

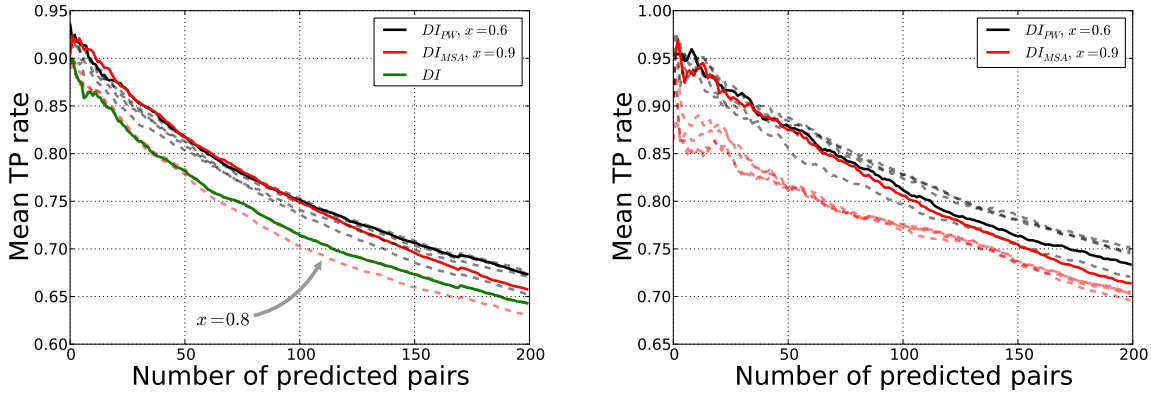
Morcos et al. [95] supplied a list of all protein families used in their study. They focused on families, where more than 90% of the family sequences belong to bacterial organisms, but they also used a dataset consisting of primarily eukaryotic proteins. We retrieved all families except for families which have been removed in the meantime, families without any atomic structure, or families having MSAs too large for our computational resources. In contrast to Morcos et al. [95] we retrieved all 3D structures for each family listed on Protein families (Pfam)[4] from the PDB. If a listed PDB entry has been updated, we used the superseding structure. The final dataset consisted of 124 bacterial (table 5.3) and 22 eukaryotic Pfam protein families (table 5.1).
Furthermore, we searched Pfam for viral protein families with at least 1000 sequences and one PDB structure. We found 18 families matching these requirements listed in the appendix (table 5.2).
The MSAs of Pfam families are based on a Hidden Markov Model (HMM) implemented with HMMER3[5] [28]. MSAs from Pfam contain upper case and lower case characters, where the latter represents residues emitted from the Insert state (I). To our knowledge, Morcos et al. [95] used only the standard amino acids as upper case characters and treated anything else as gaps.

[1] http://evfold.org
[2] http://www.python.org/
[3] https://code.google.com/p/epitopsy/
[4] http://pfam.sanger.ac.uk
[5] http://hmmer.org

3.3.3 Testing the new re-weighting algorithm

For all protein families, we calculated DI_{MSA} and DI_{PW} for re-weighting thresholds x of values $0.5, 0.6, 0.7, 0.8$, and 0.9. Then we calculated for each protein family the mean true positive (TP) rate as a function of the number of top-ranked contacts sorted decreasingly with respect to their DI value. If one or both of the amino acids which constitute a DI pair, are not crystallized in any structure, the DI pair is not included in the analysis because it cannot be evaluated. Therefore, the next pair with a lower DI value is included. A DI pair of columns i and j is a true positive if the minimal distance in any of the Pfam listed structures is $< 8\,\text{Å}$ and the involved residues are separated by at least five residues (> 5) along the sequence of amino acids. The reason for using all Pfam listed structures is that we have a better sampling of the protein, whereas limiting the analysis to a few selected structures introduces a bias in the analysis and reduces the number of sampled conformations and thus may lead to more false negatives, especially in the case of highly dynamical proteins. The distance cutoff of 8 Å has been suggested by Morcos et al. [95] because it includes all relevant physical interaction mechanisms like hydrogen bonds, salt bridges, van der Waals forces, etc. Further, they suggested the removal of pairs with a sequence separation of at least five residues because these residues are prone to coevolve due to their physical proximity and are therefore trivial predictions.

The performance of DI_{MSA} and DI_{PW} for different re-weighting thresholds is shown in figures 3.4a (bacterial dataset) and 3.4b (eukaryotic dataset). For comparison, the unweighted performance of DI is also plotted for the bacterial dataset (figure 3.4a). Comparing the best performing thresholds of DI_{MSA} and DI_{PW} shows that they perform equally well for the plotted number of predicted pairs. We assessed the performance of the different re-weighting schemes by calculating the area under the TP-curve. Interestingly, for DI_{MSA} we found the highest area under the TP-curve for $x = 0.9$, whereas Procaccini et al. [110] reported that varying the threshold from 70 % to 90 % had no significant effect. We observed that $x = 0.8$, which was used by several studies [47, 88, 95, 110], has the lowest performance and the performance is also worse than the performance of the unweighted DI. Analyzing the performance of DI_{MSA} for different thresholds shows that the performance of DI_{MSA} has a strong dependence on re-weighting. In contrast to DI_{MSA}, DI_{PW} shows a weak dependence on the value of the re-weighting threshold. For DI_{PW}, we found that $x = 0.6$ maximizes the area under the TP-curve for bacterial



(a) **(b)**

Figure 3.4: Bacterial and eukaryotic DI performance - Performance comparison of the two re-weightings DI_{PW} (black) and DI_{MSA} (red) on the bacterial dataset (left) and on the eukaryotic dataset (right). The best performing thresholds of the bacterial dataset ($x_{PW} = 0.6$ and $x_{MSA} = 0.9$) are shown in both plots as solid lines, the other thresholds are shown as dashed lines. The performance of unweighted DI is also plotted for the bacterial dataset (green).

proteins, whereas for eukaryotic proteins, the area had its maximum at $x = 0.8$. The area under the curve was calculated by summing up all TP-rates of the first 200 DI pairs sorted decreasingly with respect to their DI value and a sequence separation of at least five residues. The values of the TP-areas are listed in the appendix in tables 5.4 (bacterial), 5.5 (eukaryotic) and 5.6 (viral proteins).

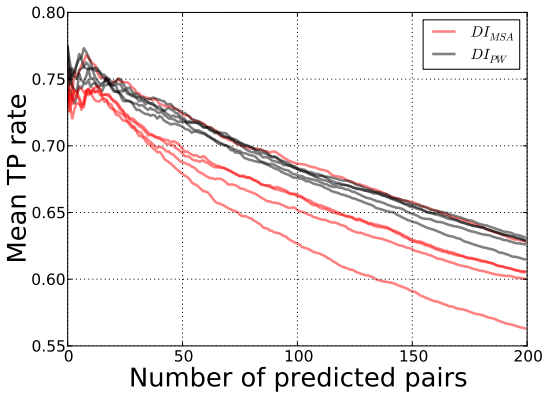


Figure 3.5: Bacterial DI performance using upper and lower case letters - Performance comparison of the two re-weightings DI_{PW} (black) and DI_{MSA} (red) on the bacterial dataset. For this dataset, all MSAs have been transformed to upper case letters, which leads to a reduced overall performance drop from a mean TP-rate of around 0.9 (only upper case) to around 0.75 (lower and upper case letters).

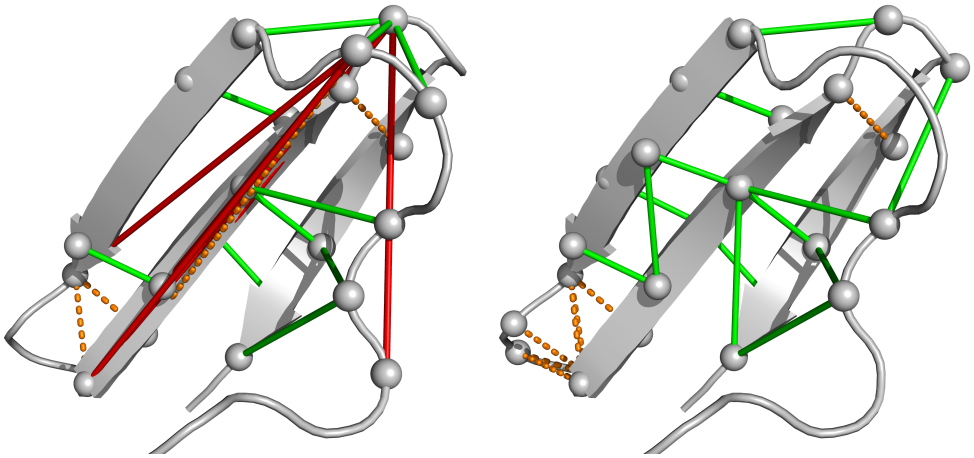
The performance of DI_{MSA} lies in the same range as reported by Morcos et al. [95], suggesting that our assumption is valid that Morcos et al. [95] treated anything but the standard amino acids in upper case letters as gaps. We also investigated the performance of DI using lower case and upper case characters for the analysis. Instead of treating anything but the standard amino acids as gaps, we removed sequences containing non-standard amino acids and transformed all characters to upper case letters, which is a more realistic application. The resulting performance for the bacterial dataset is shown in figure 3.5. The performance drops from a mean TP-rate of around 0.9 (only upper case) to around 0.75 (lower and upper case letters). Residues emitted from an insertion state are probably located in loop regions. These regions are not only characterized by structural diversity but also by a diversity on the sequence level [74]. It seems that the diversity on the sequence level introduces noise, which leads to a worse performance. Nevertheless, in

both analyses DI_{PW} has less dependence on the re-weighting threshold and has a higher TP-rate than DI_{MSA}.

We used the WD40 repeat domain (Pfam ID PF00400) as an example to illustrate the impact of the two re-weighting strategies. Therefore, we used the MSA containing both lower and upper case letters. WD40 is a structural motif found in all eukaryotes. Its function ranges from signal transduction and transcription regulation to cell cycle control, autophagy and apoptosis [75, 127]. A property of the WD40 family are the diverse sequences resulting in a low average sequence identity of 23 % leading to the inclusion of many gaps in the MSA. Figure 3.6 shows the top 20 DI pairs, with a minimum sequence separation of five, mapped on the WD40 crystal structure with the highest resolution (PDB 1yfq [144]) by different colors. For illustration purposes, a table containing the 20 DI pairs of DI_{PW} is listed in the appendix (table 5.7). We applied the best performing thresholds for the analysis ($x_{PW} = 0.6$ and $x_{MSA} = 0.9$). Green lines represent contacts fulfilled in this structure, whereas dashed orange lines denote contacts found in another WD40 Pfam listed structure. Red lines represent DI pairs which are not in contact. All predicted DI_{PW} pairs shown in figure 3.6b are in contact, indicated by green or orange coloring. In contrast, several of the predicted DI_{MSA} pairs are not in contact as illustrated by red lines (figure 3.6b).

Morcos et al. [95] also analyzed the distribution of the minimal intra-domain distances of the top 10 (20 and 30) pairs. They found a bimodal distribution with two frequency peaks around 3−5 Å and 7−8 Å. We extended the analysis to the top 200 pairs for each structure in each family and also compared the distributions of DI_{PW} and DI_{MSA} using the best performing identity thresholds ($x_{PW} = 0.6$ and $x_{MSA} = 0.9$). The resulting distance distribution of the minimal intra-domain distances between two amino acids is shown in figure 3.7a.

In agreement with Morcos et al. [95] we observed a bimodal distribution of intra-domain distances. Morcos et al. [95] related the first peak to short-range interactions like hydrogen bonds and van der Waals interactions and the second peak to long-range interactions (e.g. water mediated contacts). The height of the first peak is the same for both re-weightings. In contrast, the height of the second peak differs between the two methods and is higher for DI_{PW}. Thus, DI_{PW} enhances the performance by identifying pairs belonging to the second peak.



(a) DI_{MSA} **(b)** DI_{PW}

Figure 3.6: DI_{MSA} **vs** DI_{PW} **on WD40** - Comparison of the top 20 predicted DI pairs using DI_{MSA} (left, $x = 0.9$) and DI_{PW} (right, $x = 0.6$) on the WD40 protein family with a sequence separation of at least five residues and sorted decreasingly with respect to their DI value. Pairs are illustrated on PDB structure 1yfq as green lines if their minimal distance in this structure is $< 8\,\text{Å}$, as orange lines if their minimal distance in any other structure is $< 8\,\text{Å}$, and as red lines otherwise. For illustrative purposes, the C_α atoms are shown as spheres.

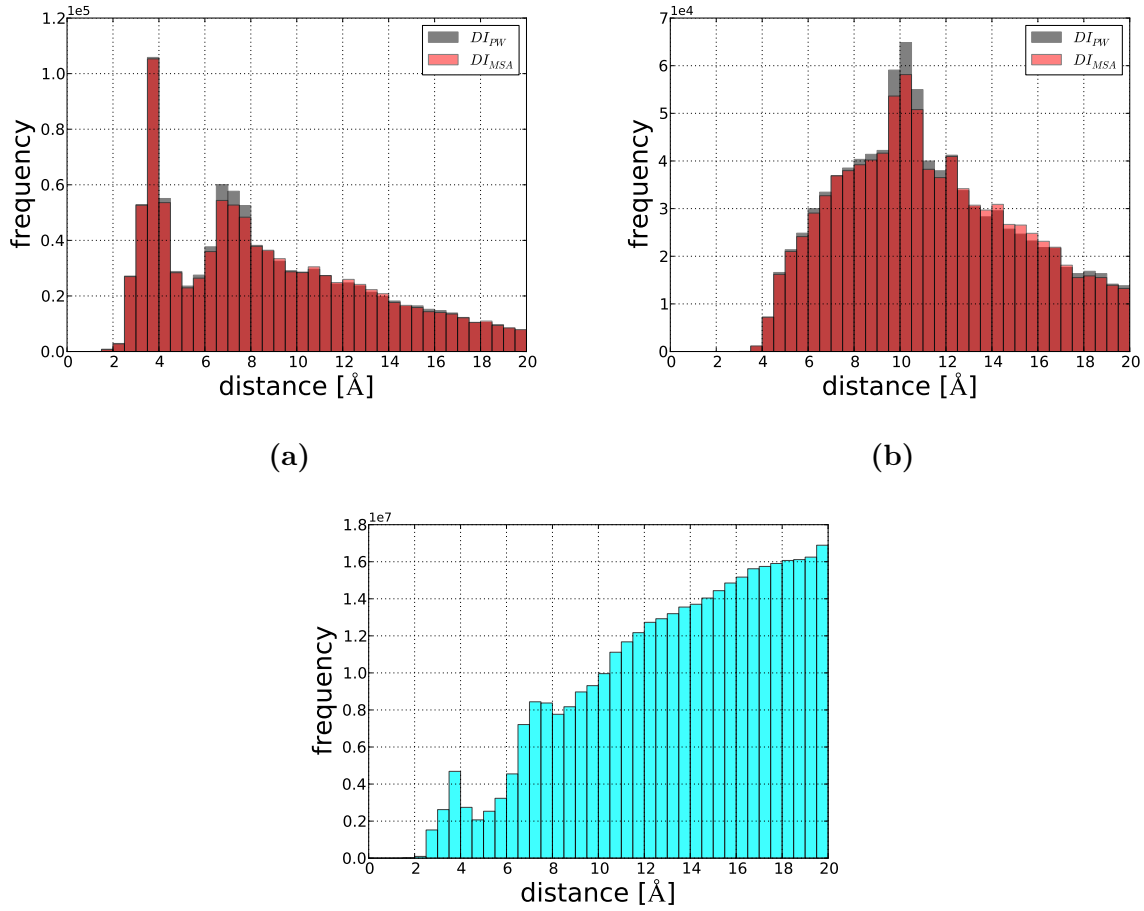


Figure 3.7: Pairwise distance distributions - Intra-domain minimal distance distribution of the top 200 DI pairs (a), intra-domain C_α distance distribution of the top 200 DI pairs (b) with DI_{PW} colored gray and DI_{MSA} colored red. The distance distribution of all residue pairs from 22 832 structures containing only proteins with a sequence length of at least 200 amino acids is illustrated in figure (c).

Additionally, we also analyzed the intra-domain C_α-distances for the top 200 pairs shown in figure 3.7b. We did not observe two peaks in the case of the intra-domain C_α-distances, suggesting that the two peaks in the previous analysis originate from side chain interactions. Morcos et al. [95] further reported that the bimodal distribution is a characteristic feature of the DI results, and it is not observed in the background distribution of all residue pairs. Therefore, we retrieved all structures from the PDB containing only protein chains with a chain length of at least 200 amino acids and a resolution between 0 and 2 Å. We limited the analysis to structures with more than 200 amino acids because small proteins may have introduced a bias for lower intra-domain distances. In the end, 22 832 structures matched this query. For each structure, we calculated all intra-domain distances. Afterwards, we calculated a histogram over all intra-domain distances ranging from 0 to 25 Å with a bin width of 0.5 Å. The histogram is shown in figure 3.7c. In contrast to the findings of Morcos et al. [95], we also found the two characteristic peaks among all amino acid pairs. Thus, the two peaks are not characteristic for the DI results. Nevertheless, it is remarkable that the DI method predominantly identifies residue pairs with a distance belonging to one of the two frequency peaks.

As already explained in the theory of DI (refer to section 3.2.2), DI is in principle MI on an isolated two-site model. Because of the connection to MI we also studied the effect of re-weighting upon the performance of MI. We used the same bacterial dataset (only upper case letters) and calculated the TP-rates for MI. The TP-curve is illustrated in figure 3.8 in combination with unweighted DI and MI. Re-weighting also enhances the performance of MI. However, the DI method remains superior of MI, despite the performance improvement.

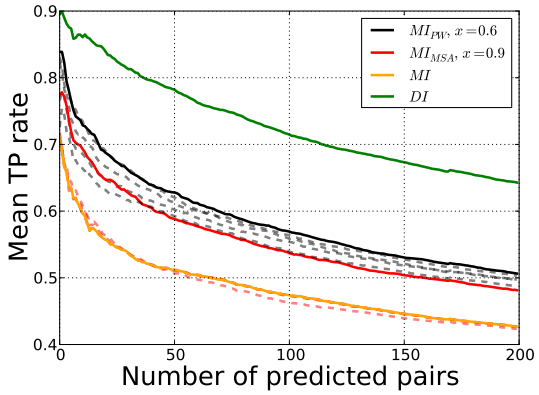


Figure 3.8: Performance comparison of the two re-weightings MI_{PW} (black) and MI_{MSA} (red) on the bacterial dataset - The best performing thresholds of the bacterial dataset ($x_{PW} = 0.6$ and $x_{MSA} = 0.9$) are shown as solid lines, the other thresholds are shown as dashed lines. Additionally, unweighted DI (green) and MI (orange) is plotted for comparison.

3.3.4 Analyzing Direct Information on viral proteins

Because Morcos et al. [95] focused on bacterial and eukaryotic proteins, the question is whether DI performs equally well on viral proteins. In contrast to bacterial organisms, viruses cannot replicate by themselves but rely on living host cells. The dependence on a host cell might constrain the evolution of viral proteins. This difference could impact the MSA and hence result in a different performance of DI.

We performed the same analysis as described in the previous subsection 3.3.3. Figure 3.9 illustrates the performance of DI_{MSA} and DI_{PW} for viral proteins. Again, the performance of DI_{MSA} and DI_{PW} is similar, but overall worse than on bacterial and eukaryotic proteins. For example, the TP value of the first DI pair droped from around 0.93 for bacterial proteins to around 0.6 in the case of viral proteins. Here, for DI_{MSA} we also observed a strong dependence on the re-weighting threshold, whereas DI_{PW} did not vary much for different thresholds, except for $x = 0.5$. This threshold refers to the lowest gray dashed line in figure 3.9.

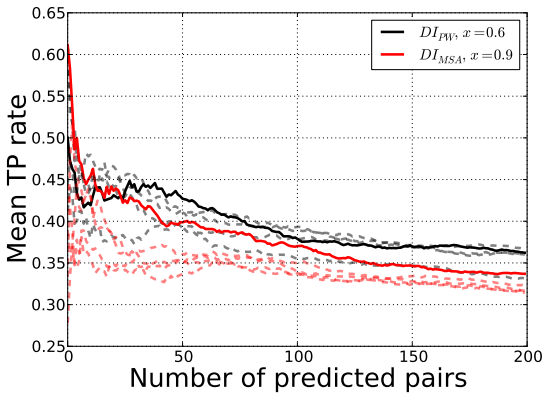


Figure 3.9: Performance comparison of the two re-weightings DI_{PW} (black) and DI_{MSA} (red) on the viral dataset - The best performing thresholds of the bacterial dataset ($x_{PW} = 0.6$ and $x_{MSA} = 0.9$) are shown as solid lines, the other thresholds are shown as dashed lines.

We reasoned that the performance drop on viral proteins is caused by the sampled sequences in the MSAs of the viral families. Therefore, we calculated histograms of the pairwise sequence identities using ten bins for each protein family. The sequence identities

were calculated according to the algorithm used for DI_{PW} as described in section 3.3. The frequencies of the histograms were normalized to one and are referred to as f_{norm}. In figure 3.10, the results are illustrated as boxplots for each bin and for bacterial, eukaryotic and viral proteins. The identities of the bacterial and eukaryotic proteins had peaks for identities ranging from 0.1 to 0.3. In contrast, the distribution for viral proteins showed no peak at all and is almost uniformly distributed. Therefore, the identities of viral proteins were in general higher. The distribution of the viral proteins could also explain why the identity threshold $x = 0.5$ has the lowest performance for DI_{PW}. Viral proteins are almost uniformly distributed and therefore half of all sequences are down-weighted for $x = 0.5$.

The explanation for the differing distributions was not part of this work. However, there are at least two possibilities. First, the viral dataset could be prone to sampling biases, i.e. sequences were mainly taken from the same species. Second, because viral proteins have no metabolism and rely on host cells for their reproduction, viral sequences might have higher constraints regarding amino acid substitutions resulting in overall similar sequences.

The distributions of the identities helped to explain the performance on bacterial, eukaryotic and viral proteins. DI requires well-sampled MSAs for the identification of interacting residues. The lower the identity of sequences, the more information they carry about the evolutionary constraints acting upon each amino acid. In the case of viral proteins the sequences were too similar and therefore the performance of DI is lower than on the datasets of bacterial and eukaryotic proteins, where the sequences had lower identity among themselves.

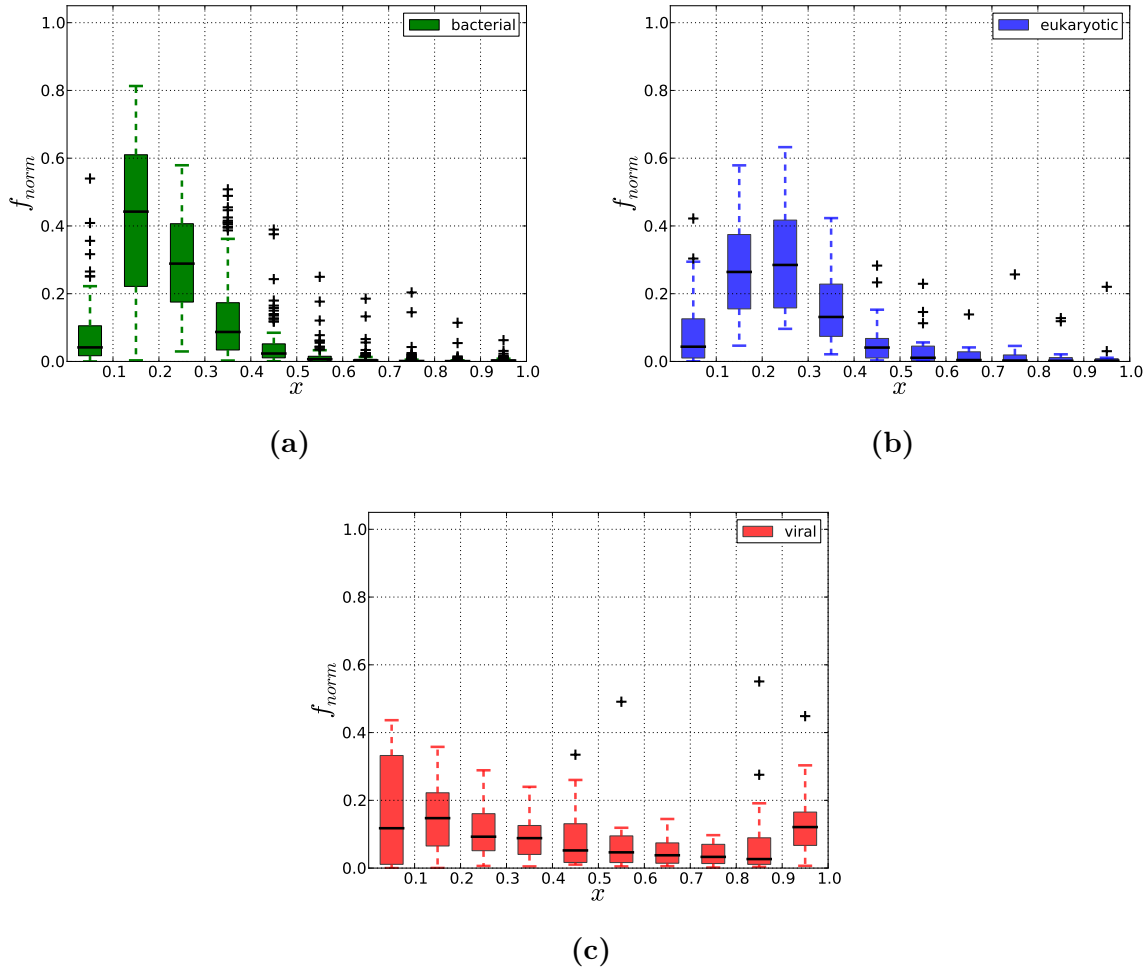


Figure 3.10: Averaged distribution of the sequence identities - Averaged distribution of the sequence identities for the bacterial (a), eukaryotic (b) and viral proteins (c). The sequence identity ranges from 0 (the two sequences have no identical amino acids at any position) to 1 (the sequences are identical).

3.4 HIV-1 envelope protein

The Human Immunodeficiency Virus-1 Envelope Protein (HIV-1 Env) is encoded in the env gene. It is first synthesized as a single entity known as Glycoprotein 160 (GP160). Before the envelope protein is presented on the viral surface, enzymes in the endoplasmatic reticulum cleave GP160 into Glycoprotein 120 (GP120) and Glycoprotein 41 (GP41). Both proteins form a non-covalently bound heterodimer. On the viral surface, three heterodimers form the final envelope spike [133, 154], which mediates the entry of HIV into the host cell. The entry process is illustrated in figure 3.11.

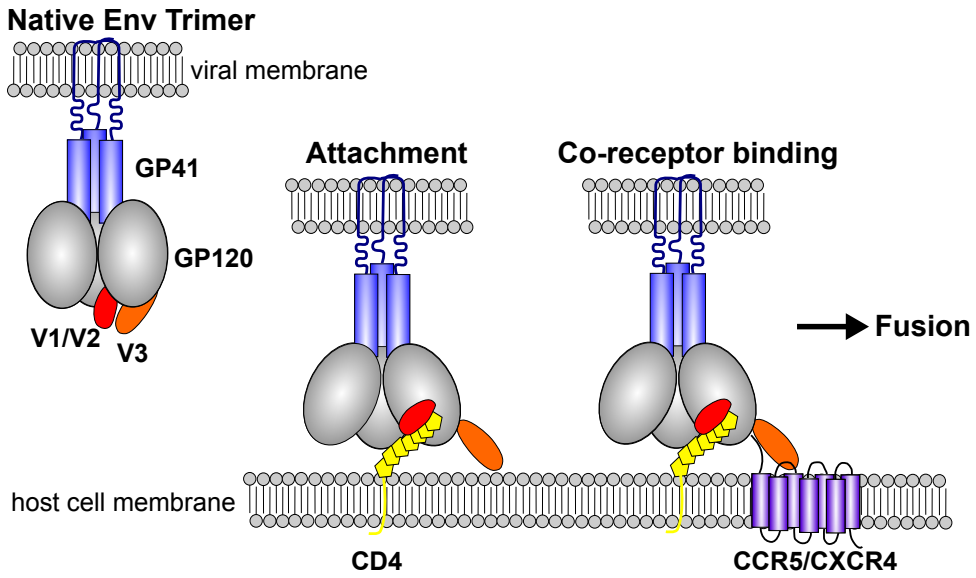


Figure 3.11: HIV cell entry - In the native Env Trimer the V1/V2 (red) form the apex of the mushroom-shaped Env trimer [79] and the V3 loop (orange) is masked by V1/V2 [80, 117]. GP41 is located in the viral membrane and GP120 is exposed to the environment. Binding of Env to CD4 (yellow) is accompanied by a conformational rearrangement, resulting in the exposure of the virus and the fusion with the host cell.

The first step of the entry is the binding of GP120 to the host cell surface receptor Cluster of Differentiation 4 Receptor (CD4), which leads to a conformational rearrangement and as a consequence V3 points away from the core of GP120, exposing the co-receptor binding

site at the stem of V3. In the next step GP120 binds to C-C Chemokine Receptor Type 5 (CCR5) or C-X-C Chemokine Receptor Type 4 (CXCR4), reviewed by Klasse [62]. This interaction induces other rearrangements and triggers the fusion of virus and host cell.
The conformational dynamics during the entry are not yet fully understood. So far a number of low-resolution electron-microscopy structures of the Env complex for different states of the entry are available [48, 79, 87, 133, 143, 147]. Although high resolution structures of GP41 [13, 140] have been solved, the understanding of its dynamics is far from complete. As a membrane protein, it is even more difficult to produce high quality structures. In the case of GP120, several atomic-level structures in complex with CD4 and/or ligands/antibodies have been reported [49, 50, 68, 69, 102]. Recently, Kwon et al. [67] solved an atomic-level unliganded structure of GP120 core. Unfortunately, none of these atomic-level structures have been solved including all variable loops, due to their high flexibility. The conformation of V3 in the CD4 bound state has been solved by Huang et al. [49, 50]. Mao et al. [86] published a 6 Å cryo-electron microscopy structure of a membrane bound HIV-1 Env trimer, although its validity is still under discussion [18] and therefore their structure and findings are not considered here.
Several studies found interactions between V1/V2 and V3 [42, 48, 80, 87, 117]. These interactions are described as a mechanism of HIV to shield its co-receptor binding site at the stem of V3 from potential antibodies. However, there are conflicting hypotheses whether it is an intra- or inter-GP120 interaction [80, 117].
There have already been studies using MSAs for the inference of structural and functional properties of GP120 [64, 138]. In this study, we used DI for the analysis of HIV-1 Env to gain further understanding of the GP120-GP41 complex.

3.4.1 Data extraction and preparation

We retrieved all available HIV-1 Env sequences from the Los Alamos database[1] using only the option "one sequence per patient" to filter sequences belonging to the same patient to prevent sampling biases. Furthermore, all sequences containing non-standard amino acids were removed from the dataset, resulting in an MSA consisting of 4844 sequences. The dataset consists primarily of HIV subtypes B and C, which together account for over 60 % of all sequences. A list of all subtypes included in the MSA is located in the appendix in table 5.8.

[1] http://www.hiv.lanl.gov/

3.4.2 Choosing a re-weighting threshold

In subsection 3.3.4 we found that the performance of DI on viral proteins is lower than on bacterial and eukaryotic proteins. We attributed the low performance to the distribution of pairwise sequence identities of the MSA. In figure 3.12, a histogram over the pairwise sequence identities for HIV-1 Env is shown. The histogram of the identities has its maximum value for the bin including identities between 0.7 and 0.8. The shift to high identities compared with the ones found for bacterial and eukaryotic proteins stems from the selection of sequences. Here, we included only sequences sampled from the human organism, whereas protein families of the bacterial and eukaryotic dataset contained sequences from several hundred species. For GP120 it is known that it consists of conserved and variable regions [128]. The conserved regions are located in the core, whereas the variable regions are mostly located at the surface and in the loops (V1-V5).

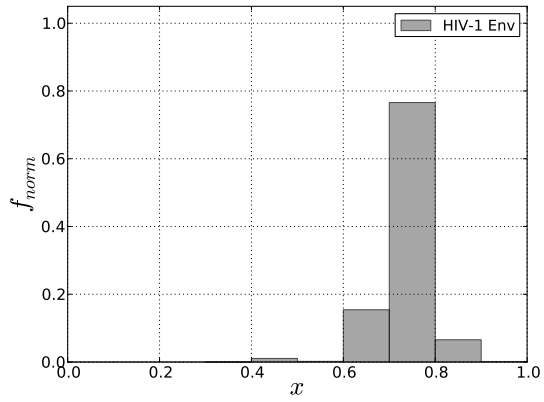


Figure 3.12: Sequence identity distribution of HIV-1 Env - Normalized histogram of the pairwise sequence identities of the HIV-1 Env MSA. Most of the sequences have an identity in the range of 0.6 to 0.7.

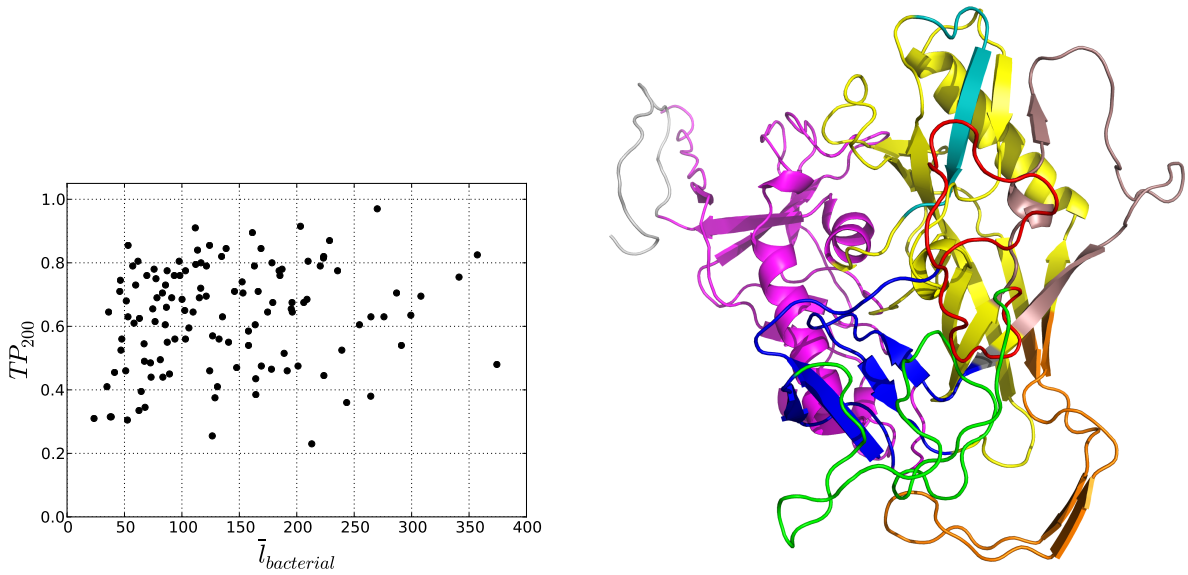
Re-weighting has been introduced for the correction of sampling biases. Although we found that the performance of DI_{PW} is less dependent on the choice of the re-weighting threshold, one has to be careful not to choose a threshold that is lower than the peak of the distribution of the pairwise sequence identities because otherwise almost all sequences are down-weighted. For HIV-1 Env, we found that most of the sequences have an identity ranging from 0.7 to 0.8. In the following, we therefore applied a threshold of $x = 0.8$ for the re-weighting.

3.4.3 Modeling full GP120

For GP120, no crystal structure is available containing the whole sequence. We employed Modeller [119] for the construction of a full GP120 structure. Therefore, we used two representative structures of GP120 from the PDB, 2qad [49] and 3jwd [102].
A problematic step in our analysis was the choice of the number of included DI pairs n_{DIs}. The inclusion of too many DI pairs may result in a high false positive rate. However, by including only a few pairs, we might miss interesting DI pairs. The performance of the bacterial dataset using all characters of the MSA (figure 3.5) was our reference benchmark, which contained 124 well-sampled protein families. The mean TP rate of the 200th prediction is $TP_{200} = 0.63$. However, HIV-1 Env has an average protein length $\bar{l}_{HIV-1Env} = 857$, whereas the longest average protein length of the bacterial dataset is $\bar{l}_{max} = 374$. We analyzed if a correlation between the average length and the TP rate of the 200th prediction exists. Pearson's correlation coefficient $\rho = 0.163$ for the data shown in figure 3.13a suggests that no strong correlation exists. Therefore, we also used the first 200 predicted DI pairs sorted descendingly with respect to their DI value and a sequence separation of at least five residues.
The distance restraints were implemented using an upper bound potential with a mean of 11 Å between the two C_α atoms of the DI pair and a standard deviation of 1 Å. We applied this distance cutoff after evaluating the C_α distances of the top-ranked 200 DI pairs of the bacterial and eukaryotic dataset. The distribution is shown in subsection 3.3.3 in figure 3.7c. Further restraints were added for the cysteine bridges located in V1/V2 between residues 126-196, 131-157 and 119-205. During the modeling process, we set the *md_level* parameter to very slow, which increases the number of steps at each temperature and number of temperatures to use during the simulated annealing. The optimization was repeated 20 times.
The conformation of the resulting GP120 structure based on these restraints is only used for illustration purposes. Evolutionary constraints arise due to the interaction of residues in spatial proximity. V1/V2 have a dynamical role during the entry of HIV into the host cell as reported by several studies, e.g. [42, 79, 133]. Therefore, it is plausible that the sequence region of V1/V2 contains evolutionary constraints based on different conformations. Structure refinement based on DI constraints is not able to distinguish

between constraints from different conformations and thus the resulting structure is a mixture of all states.

One example is the V1/V2 - V3 interaction. In the CD4 bound conformation V3 is extended away from the protein and there is no interaction between V1/V2 and V3, whereas in the native state, there is evidence for V1/V2 - V3 interaction [80, 117]. Within the top-ranked DI pairs we also found V1/V2 - V3 interaction. Thus, any model based on our DI predictions will be more similar to the native GP120. The resulting model is shown in figure 3.13b and analyzed in the following subsection.



(a) **(b)**

Figure 3.13: DI constrained model of GP120 structure. - (a) illustrates the TP rate for the 200th prediction in dependence of the average length of the protein family. (b) shows the full GP120 model colored according to the domain definition of table 3.1.

V1	V2	V3	V4	V5	ID	OD	BS
132-156	157-196	296-339	385-418	461-471	33-118 206-252 476-511	253-397 410-419 437-475	118-205 421-437
red	green	orange	dirtyviolet	teal	magenta	yellow	blue

Table 3.1: Overview of the domain range and applied color scheme - The domain range is based on the definition by Kwon et al. [67] and extended for full GP120. The names of the colors refer to the pymol [123] colors.

3.4.4 Analysis of the predicted DI-pairs for HIV-1 Env

We limited the analysis of HIV-1 Env to three domains of HIV-1 Env, namely V1/V2, V3 and GP41. For each domain we analyzed the intra- and inter-domain couplings, except for GP41, where no structure is available, so we focused on the intra-domain residues. The results are illustrated in figure 3.14. A table containing all 200 DI pairs is shown in the appendix in table 5.9.

The conformation of V1/V2 in our model (figure 3.13) is solely based on 27 DI-based constraints for 63 residues because no further structural restraints are available. Thus, the conformation of the loop is disordered and not easy to interpret. Therefore, all intra-domain DI pairs of V1/V2 are illustrated in a schematic representation in figure 3.14a. We assume that the three cysteine bridges exist regardless of the actual V1/V2 conformation. Two structures of a scaffolded V1/V2 have been published in the PDB, with one structure missing parts in V1 and V2 [103] and one structure only missing parts in V2 [92].

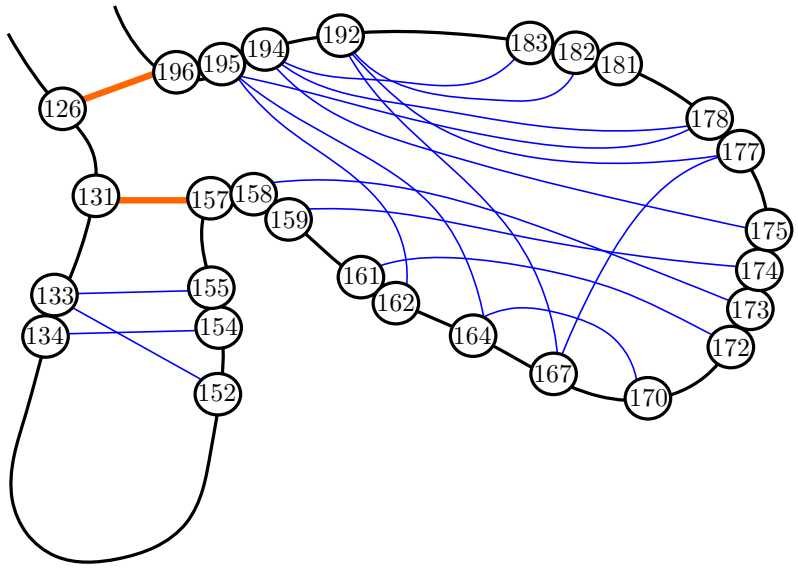
We found three predicted DI couplets located inside V1, which are almost parallel to the cysteine bridge 131-157. The placement of these couplets suggests an extended loop conformation, which is in agreement with the scaffolded V1 structure found by McLellan et al. [92]. In V2, we found 14 intra-domain contacts. Interestingly, we found a clear pattern amongst the intra-domain contacts. The V2 loop can be divided into four parts ranging from 157-166, 167-176, 177-186 and 187-196. We found DI pairs connecting the first and the second part in reversed order. The same pattern can be observed for parts three and four. The pattern can be interpreted as a four-stranded antiparallel beta sheet structure. In the scaffolded V1/V2 structures McLellan et al. [92] and Pancera et al. [103] also found that V1/V2 assumes a beta sheet conformation, although some parts of V2 could not be resolved. For V2, we also found two couplings that range from part one to part four of V2. However, these couplings do not necessarily contradict the four-stranded antiparallel structure because these pairs could still be in contact depending on the kind of interaction and involved residues. Another possibility is that during the life time of GP120 the V2 loop assumes an extended conformation, where these DI pairs form structurally or functionally important interactions.

Furthermore, we found three inter-domain contacts connecting V1/V2 with V3. This loop interaction has already been proposed in the literature [80, 117]. Assuming the

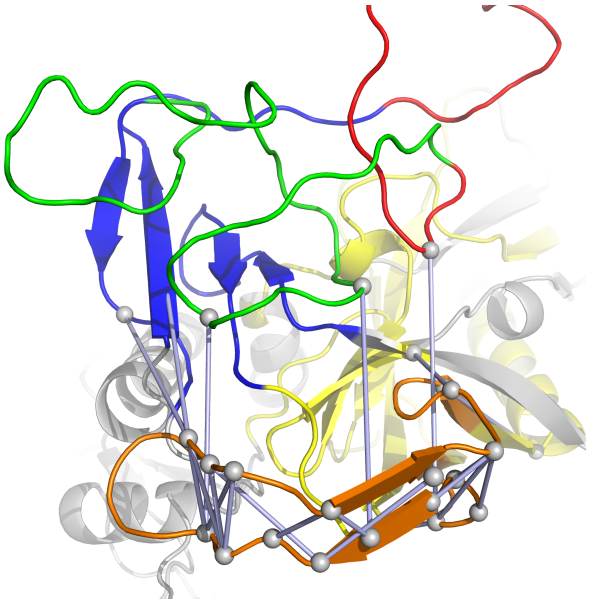
conformation of V2 is an antiparallel beta sheet, the DI pairs connecting V1/V2 and V3 are also in the correct order, which can be seen in figure 3.14b. The residue position 154 located in V1 is in contact with residue 300 of V3. Residue 172 located in V2 is closer to 154 and is in contact with residue 305 of V3. The last residue is 167 in V2, which is in contact with residue 309 of V3. This coupling was also part of our previous analysis [136], where we used a small set of succinct signature patterns to distinguishes Chinese and non-Chinese HIV-1 genomes. In our study, we also proposed a DI pair between 170 and 317. For the study, we used DI_{MSA}, although the contact is not part of our current analysis, but it is nevertheless in agreement with our current analysis.

The DI pairs involving V3 are illustrated in figure 3.14b. We found 13 intra-domain pairs, which are all parallel to the cysteine bridge between residues 296 and 331. Further, we found two pairs contacting the residue 315 at the tip of V3 to residues 121 and 202 located in the bridging sheet. The model suggests that V3 maintains its beta sheet structure, but is docked to the core of GP120 and V1/V2 are in close spatial proximity of V3, with the tip of V3 located at the bridging sheet.

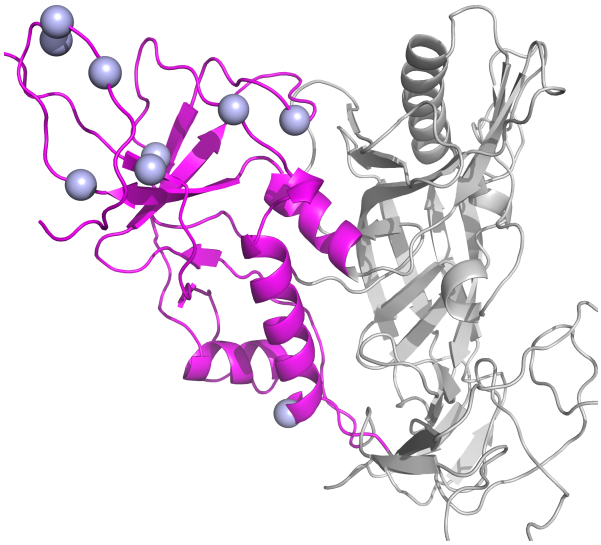
For GP41, we found nine inter-domain DI pairs between GP41 and GP120. Because no structure of the GP120-GP41 complex exists, we illustrated only the DI partners located in GP120. The results are shown in figure 3.14c. All interaction partners are located in the inner domain. Pancera et al. [102] have gathered GP41 interaction sites located in GP120 based on mutagenesis experiments in their study. Three of our interaction partners (residues 92, 499 and 500) have been annotated as in contact with GP41. Further four residues (46, 236, 492 and 502) are direct sequence neighbors of an annotated interaction partner with GP41, residue 244 is two residues away from an annotated interaction partner. Only residue 114, which is located at the end of the second alpha helix has not been annotated by Pancera et al. [102]. Either this residue is a false positive or it might interact with GP41 during the fusion of HIV into the cell.



(a) V1/V2



(b) V3



(c) GP41

Figure 3.14: DI analysis of three domains of HIV-1 Env - (a) Schematic representation of intra-domain DI pairs of V2, colored according to table 3.1. (b) DI pairs which have at least one residue located in V3. (c) depicts the interaction partners of GP41 in GP120 as spheres.

3.5 Summary and outlook

The introduced modification to the re-weighting scheme improves the performance of DI. By calculating pairwise sequence identities for all sequence pairs instead of using the complete rows of the MSA, the performance is also more independent of the threshold used for re-weighting. Re-weighting also enhances the performance of MI, although the increase is not enough to reach the performance of DI. Future experiments should address the possibility of using re-weighted MI for structure prediction. MI is not limited by the available memory, whereas DI requires large amounts of memory for the storage of single site and pair frequencies. With the ever-growing number of available sequences due to the advances in sequencing methods [81], DI will be a useful tool in the analysis of structures and functions of proteins.

In the presented example of HIV-1 Env we could show that DI is capable of assisting in the analysis of complex bimolecular structures. The loops of GP120 have always challenged experimenters because of their flexibility. Our data supported the theory that V1/V2 interacts with V3. Furthermore, the data also suggests that V3 interacts with residues located in the bridging sheet. Together the data suggest that V2 assumes the conformation of a four-stranded antiparallel beta sheet. DI pairs located in V3 suggest that V3 always assumes the two-stranded antiparallel beta sheet conformation and is in spatial proximity of V1/V2 and that the tip of V3 is in close proximity of the bridging sheet. Further, we found that DI predictions of the interaction partners between GP41 and GP120 are also in good agreement with results from mutagenesis studies. All predictions are consistent with each other and most importantly also with reported experiments in the literature. Although we included sequences from only one species, we gained further insight into the HIV-1 Env machinery. The proposed interactions of V1/V2 and V3 can be the subject of further experiments.

4

Analysis of heparan sulfate interacting regions of proteins

Many theoretical and computational methods are available to address protein-protein interactions, including purely sequence-based methods as DI in the previous chapter. These methods usually cannot be applied to important interactions as between proteins and Glycosaminoglycans (GAGs) that dominate the extracellular space. We developed an algorithm based on the electrostatic properties of heparan sulfate (HS), a member of the GAG family, and proteins for the prediction of favorable regions of interaction on proteins. The algorithm is introduced in section 4.2.1. We have applied our model for the prediction of favorable regions around the chemokine CCL3 (section 4.3) and for proteins of the Hedgehog family (section 4.4). A detailed introduction on the proteins is given in their respective sections.

4.1 Glycosaminoglycans

GAGs are linear polysaccharides of disaccharide units consisting of an amino sugar (N-acetylglucosamine or N-acetylgalactosamine) combined with a uronic sugar (glucuronic acid or iduronic acid) or galactose [29]. GAGs are extremely heterogeneous regarding the composition of disaccharides, sulfation patterns and the total molecular weight. The properties of cell surface GAGs can also be cell-type specific and especially regulated depending on the developmental state of the organism [54, 145]. In addition, they can also be affected by a pathophysiological state, such as cancer [78].

Heparin is a member of the GAG family only produced in mast cells [29] and is released near a tissue injury to the vasculature. In heparin, the uronic acid is predominantly L-iduronic acid (IdoA) (>70%) or to lesser extent D-glucuronic acid (GlcA) [111]. The glucosamine can be either unmodified (GlcN), N-sulfonated (GlcNS), or N-acetylated (GlcNAc) [55]. Heparin is the biopolymer with the highest charge density [11]. This also limits the conformational flexibility of heparin resulting in a semi-rigid helical structure [51, 61]. The molecular weight of heparin ranges from 7 kDa to 20 kDa [29] and the average heparin disaccharide contains 2.7 sulfate groups [11].

Heparan sulfate is structurally similar but more abundant than heparin as it is synthesized in almost all cells [29]. The uronic sugar is also either IdoA or GlcA, although GlcA is more frequent than IdoA. Like heparin, the uronic acid is linked to glucosamine. In contrast to heparin, HS on average only contains 1.0 sulfate group per disaccharide, but there is a high diversity regarding the distribution of the sulfate groups, i.e. there are regions with only few sulfate groups and regions with a high sulfate content, comparable to that of heparin. It is thought that the sulfated regions are functionally significant, and the unsulfated regions are spacers between active domains [111]. The sequence of HS can also be specific regarding the interaction with proteins [37]. The degree of sulfation affects the conformational flexibility of the polysaccharide because a high content of sulfate groups leads to electrostatic repulsion of the charges and therefore reduces the conformational freedom. Chains of HS also tend to be longer than heparin with a molecular weight ranging from 10 kDa to 70 kDa [29]. HS occurs as proteoglycan (HSPG), where one or more chains are covalently bound to a core protein, on cell surfaces or in the extracellular matrix.

The characterization of heparin and HS is challenging, because of the diverse primary structure (especially for HS) and the difficulty of determining the sequence experimentally [17, 111]. Therefore, most of the experimental studies have been performed with heparin. From these studies it is known that heparin and HS show a wide range of biological functions, i.e. receptor engagement, signaling, protein function, cellular adhesion, regulation of cellular growth and proliferation, promotion of oligomerization and protection from proteolysis reviewed in [11, 44, 111].

4.2 Methods

In this section, we present our electrostatic interaction energy model. At first, we explain the theoretical background of the model followed by theoretical descriptions of the electrostatic potential and the Fast Fourier Transformation (FFT) correlation, both essential to our model. Further we also present the set-up of our calculations and scoring functions, which can be implemented in an alanine scanner to assess the contribution of individual residues to the interaction of HS with proteins. The section ends with a comparison of our model with known HS binding sites.

4.2.1 Electrostatic interaction energy model

As previously mentioned HS, especially its close relative heparin, contains a huge amount of negative charges. Therefore, the idea behind this method is that the protein-HS interaction is, for the most part, driven by electrostatic forces. The length and high flexibility of HS chains are not well suited for normal docking softwares, therefore the focus lies on identification of favorable regions on proteins, where HS encounters the protein. The final binding site may be reached after local structural rearrangements. The approach is similar to that described by Gabdoulline and Wade [36], where the electrostatic interaction energy is considered for a representative set of relative positions and orientations of a protein and a GAG ligand. Only positions without overlap between the protein and the ligand are taken into account. For the detection of overlapping positions, we employed the FFT correlation technique suggested by Katchalski-Katzir et al. [58]. The underlying theory is given in subsection 4.2.3. For each non-overlapping position the energy is given by

$$E_{estat}(\vec{r}) = \sum_{i=0}^{N_q} \phi(\vec{r}_i) q(\vec{r}_i), \tag{4.1}$$

with the electrostatic potential ϕ generated by the charges of the fixed protein, q is a charge of an HS atom, $\vec{r}_i$ is the position of that charge and N_q is the total number of charges on the HS molecule. To reduce the computational time we followed the instructions by Gabb et al. [35] and calculated the energies using FFT correlation as well. In principle, one could define the electrostatic interaction energy as the thermodynamic total energy

for each grid point:

$$\langle E(\vec{r})\rangle = \sum_{j=1}^{\Omega} \frac{E_{estat,j}(\vec{r})}{Z} \exp\left(-\frac{E_{estat,j}(\vec{r})}{k_B T}\right), \tag{4.2}$$

where Ω is the number of sampled orientations, k_B is the Boltzmann constant and T is the temperature. Z is the partition function, which is defined as:

$$Z = \sum_{\vec{r}} \sum_{j=1}^{\Omega} \exp\left(-\frac{E_{estat,j}(\vec{r})}{k_B T}\right) \tag{4.3}$$

and is in principle calculated over infinite space. Therefore, the sum $\sum_{\vec{r}}$ in the partition function contains infinite summands which cannot be computed numerically. We chose a different approach in which we compare the probabilities for a set of rotations of the ligand around the protein $p(\vec{r})$ and in solution $p(\vec{r}_{solv})$. The probability is defined as

$$p(\vec{r}) = \frac{1}{Z} \sum_{j=1}^{\Omega} \exp\left(-\frac{E_{estat,j}(\vec{r})}{k_B T}\right). \tag{4.4}$$

In an isotropic solution, all rotations should have the same energy, which equals zero. Thus, the probability in solution is given by

$$p(\vec{r}_{solv}) = \frac{\Omega}{Z}. \tag{4.5}$$

In the following, we interpret the probabilities as concentrations. For example, if the probability of the ligand close to the protein is twice the probability of the ligand in solution, we interpret this as the concentration of the "bound" state is twice as high as of the "unbound" state. Using this definition, we can define the equilibrium constant as the ratio of the probabilities $p(\vec{r})$ and $p(\vec{r}_{solv})$

$$K = \frac{p(\vec{r})}{p(\vec{r}_{solv})} = \frac{1}{\Omega} \sum_{j=1}^{\Omega} \exp\left(-\frac{E_{estat,j}(\vec{r})}{k_B T}\right). \tag{4.6}$$

This equilibrium constant K can be related to the difference in Gibbs free energy $\Delta G(solv \rightarrow prot)$ (here we only calculate the electrostatic component):

$$\Delta G(solv \rightarrow prot)(\vec{r}) = -k_B T \ln(K) = -k_B T \ln\left(\frac{1}{\Omega} \sum_{j=1}^{\Omega} \exp\left(-\frac{E_{estat,j}(\vec{r})}{k_B T}\right)\right). \tag{4.7}$$

4.2.2 On the theory of the electrostatic potential

The energy model in equation 4.1 is based on the electrostatic potential ϕ generated by the charges of the protein. The electrostatic potential can be solved numerically with the Poisson-Boltzmann equation. The formula can be derived by starting with Gauss's law, which states that the charges are the sources of the electric field,

$$\vec{\nabla} \cdot \vec{D} = \frac{\rho_{free}}{\varepsilon_0}, \tag{4.8}$$

where $\vec{\nabla} \cdot \vec{D}$ is the divergence of the electric displacement field, ρ_{free} is the free charge density and ε_0 is the vacuum permittivity. The displacement field is related to the electric field

$$\vec{D} = \varepsilon \vec{E}, \tag{4.9}$$

with permittivity ε of the material. Electrostatics imply a static system and therefore all time derivatives are zero and the Maxwell-Faraday equation is

$$\vec{\nabla} \times \vec{E} = 0, \tag{4.10}$$

which means that the electrostatic potential is conservative, and it is given by a gradient of the potential

$$\vec{E} = \vec{\nabla}\phi. \tag{4.11}$$

In the next step, the electric displacement $\vec{D}$ in equation 4.8 is replaced with equation 4.9 and 4.11, which gives Poisson's equation for an inhomogeneous medium

$$\vec{\nabla} \cdot \left(\varepsilon(\vec{r})\vec{\nabla}\phi(\vec{r})\right) = -\frac{\rho(\vec{r})}{\varepsilon_0}. \tag{4.12}$$

In addition to Poisson's equation, which solves the electrostatic potential for any charge distribution and shielding effects due to the permittivity, the Poisson-Boltzmann equation also accounts for mobile ions e_+, e_-

$$-\varepsilon_0\vec{\nabla} \cdot \left(\varepsilon(\vec{r})\vec{\nabla}\phi(\vec{r})\right) = \rho(\vec{r}) + nz_+e_+ - nz_-e_-, \tag{4.13}$$

with the valency z. In the following, the valency z is set to 1, because the effects accompanied with a valency $z > 1$ go beyond the continuum approach of the

Poisson-Boltzmann approach (e.g. [19, 60]). The ions adjust to the electrostatic potential in thermal equilibrium and their density at position $\vec{r}$ can be described by a Boltzmann distribution

$$n_{\pm}(\vec{r}) = n_{\infty} \exp\left(\frac{\pm e\phi(\vec{r})}{k_B T}\right), \tag{4.14}$$

where n_{∞} is the density for which $\phi = 0$ and k_B is the Boltzmann constant. The final Poisson-Boltzmann equation is then given by

$$-\varepsilon_0 \vec{\nabla}\left(\varepsilon(\vec{r})\vec{\nabla}\phi(\vec{r})\right) = \rho(\vec{r}) - 2en_{\infty}\sinh\left(\frac{e\phi(\vec{r})}{k_B T}\right). \tag{4.15}$$

4.2.3 On the theory of the Fast Fourier Transformation Correlation

The geometrical shape complementary of two structures s_1 and s_2 can be calculated using the FFT-Correlation. The structures are mapped onto a three-dimensional grid, where they are represented as discrete values

$$s_{1,x,y,z} = \begin{cases} 1, & \text{on the surface of the structure} \\ \rho, & \text{inside the structure} \\ 0, & \text{outside the structure} \end{cases} \tag{4.16}$$

and

$$s_{2,x,y,z} = \begin{cases} 1, & \text{on the surface of the structure} \\ \delta, & \text{inside the structure} \\ 0, & \text{outside the structure} \end{cases} \tag{4.17}$$

with grid indices x, y and z. A grid point is defined as inside the structure if its distance to any atom is less than the van der Waals radius of that atom and defined as outside if its minimal distance to any atom is larger than the van der Waals radius of that atom. The surface is defined as the layer of grid points, which separates the grid points considered as inside from the ones considered as outside of the structure. The correlation between the two structures can then be defined as

$$c_{\alpha,\beta,\gamma} = \sum_{x=1}^{N} \sum_{y=1}^{N} \sum_{z=1}^{N} s_{1,x,y,z} \cdot s_{2,x+\alpha,y+\beta,z+\gamma}, \tag{4.18}$$

where α, β and γ constitute a shift vector in units of grid spacing along each dimension. If the shift vector points to a new position, where both structures have no contact, the correlation is 0. For shift vectors which lead to contact of the two structures the correlation is either positive or negative, depending on the values of ρ and δ and the amount of overlap between the structures. Using the definition of the Discrete Fourier Transformation (DFT)

$$F_k = \sum_{x=1}^{N} f_x e^{-2\pi i(xk)/N}, \tag{4.19}$$

with the imaginary unit $i = \sqrt{-1}$, equation 4.18 can then be formulated as

$$C_{j,k,l} = S^*_{1,j,k,l} \cdot S_{2,j,k,l}, \tag{4.20}$$

where S_2^* is the complex conjugate of the DFT of s_2. The correlation can be obtained after the inverse Fourier transformation

$$c_{\alpha,\beta,\gamma} = \frac{1}{N^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} e^{2\pi i(j\alpha+k\beta+l\gamma)/N} C_{j,k,l}. \tag{4.21}$$

The highest peak in $c_{\alpha,\beta,\gamma}$ represents the shift vector, which maximizes the shape correlation of the two structures. For the search of the global optimum the calculations have to be repeated for all relative orientations of structures s_1 and s_2.

4.2.4 Implementation of the algorithm

During his diploma thesis, Dybowski [27] developed a Java framework for computational comparison of protein surfaces named "epitopsy"[1]. This project served as the basis for this analysis. The code was ported to Python[2] because of its interactive data exploration capabilities. The code was optimized by using NumPy [100] for the N-dimensional array objects and Cython [4] for the generation of fast C-Extensions. Further, we implemented the FFT using anfft[3], which interfaces to FFTW3 [34] and is up to 10 times faster than the NumPy implementation.

[1] https://code.google.com/p/epitopsy/
[2] http://www.python.org/
[3] https://code.google.com/p/anfft/

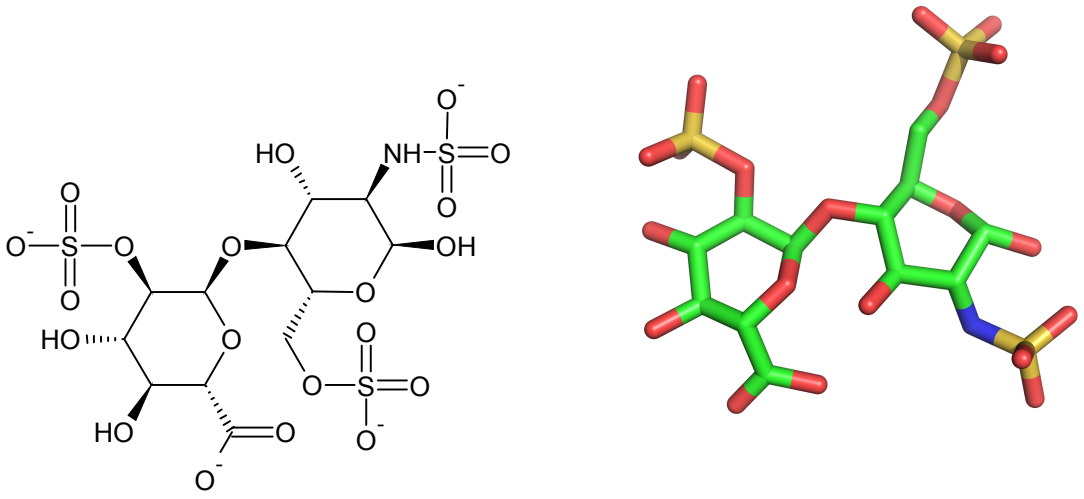
4.2.5 Set-up of the calculations

The electrostatic potential ϕ was calculated by solving the non-linear Poisson-Boltzmann equation 4.15 with APBS (Version 1.3) [1]. Radii and charges of the fixed protein were assigned with PDB2PQR (Version 1.8) [24] using the AMBER forcefield parameters. For calcium ($r_{Ca} = 1.76\,\text{Å}$, $q_{Ca} = 2\,\text{e}$) and zinc ($r_{Zn} = 1.24\,\text{Å}$, $q_{Zn} = 2\,\text{e}$) ions, radii r and charges q were added manually because they are not part of the forcefield. If not stated otherwise, all calculations used the following input parameters for APBS. The grid dimensions were chosen to cover the whole system by a grid with spacing of 0.05 nm, which is sufficient to produce reliable electrostatic potentials [45]. The dielectric constant was set to 79 outside and 2 inside the protein. Monovalent ions with a concentration of 0.15 mol/l were added to the system, and the temperature was set to 310 K. For the boundary condition, the Multiple Debye-Hückel algorithm was employed. The surfaces of the proteins were approximated by a solvent probe with a radius of 0.14 nm.
The topology and charges of the HS molecule were calculated with the PRODRG2 Server [124]. Radii were assigned with PDB2PQR (Version 1.8) [24] using the AMBER forcefield parameters.
The orientation-averaged electrostatic energies from equation 4.2 were calculated for 150 different orientations of the HS molecule, calculated with a golden section spiral algorithm[1]. For the geometrical matching, the interior of the fixed structure was set to -15, and the interior of the rotated structure was set to 1, according to the values suggested by Katchalski-Katzir et al. [58].
All energies are in units of k_BT. For a given temperature T, particles in a solvent have an energy $\approx k_BT$, thus energies $|E| > k_BT$ are therefore above the thermal noise and strong enough to induce electrostatically-driven dynamics.

[1] http://www.softimageblog.com/archives/115

4.2.6 Selection of an HS probe

For the computational screening of favorable interaction sites of HSs around proteins a disaccharide consisting of fully-sulfated IdoA linked to GlcNS was constructed in cooperation with Jan Taubenheim, Research Group Bioinformatics, University of Duisburg-Essen and the software tool PRODRG [124]. The choice of the fully sulfated disaccharide is motivated by the assumption that the sulfate is the active constituent of HS [37]. The algorithm described in section 4.2.1 uses the shape of a single structure for the exclusion of overlap between the protein and the ligand, and hence the assumption is that the impact of the difference between the isomers IdoA and GlcA can be neglected. The isomers also have a different charge distribution, but because we also neglect any sampling of the HS probe, we can also neglect the difference in charge distribution caused by the isomer. In theory, the probe could also consist of more than two disaccharide units, but this would require sampling of internal conformational degrees of freedom, because for longer chains the error due to insufficient sampling has a profound effect. Figure 4.1a shows the Natta projection and figure 4.1b illustrates the resulting atomic structure of the molecular probe.



(a) (b)

Figure 4.1: HS-fragment used as a probe in electrostatics calculations - Left: Natta projection of the HS-fragment. Right: Resulting atomic structure of the HS-fragment.

4.2.7 Functions for the analysis of the interaction between HS and proteins

An easy and intuitive way to analyze the energy landscape around proteins is to simply visualize isosurfaces at a certain threshold. However, these visualizations cannot be quantified easily. Therefore, we devised four heuristic functions and one based on a physical interpretation to assess the interaction of HS and proteins. The input for the functions is a matrix containing the difference in Gibbs free energies $\Delta G(\vec{r})$ for each grid point. For the analysis we can define a surface S around the protein and a volume V which contains all grid points of the matrix except the ones on the surface and inside the surface S. First, we counted the number of grid points A_+ on the surface S for which $\Delta G(solv \rightarrow prot) < -1\,\mathrm{k_B T}$. Second, we weighted the grid points on the surface with their corresponding energy values referred to as wA_+ in the following. For the third function, we counted the number of grid points V_+ contained in the volume V which have an energy $\Delta G(solv \rightarrow prot) < -1\,\mathrm{k_B T}$. Fourth, we also weighted these grid points with their corresponding energy and refer to this value as wV_+. The idea was that the surface scores correlate with the binding strength of HS to the protein, whereas the volume scores correlate with the long-range attraction of HS to the protein.

For proteins which contain a large number of charges and thus have a high potential the volume scores V_+ and wV_+ can be prone to error if the size of the box is too small and the values at the borders of the box are $< -1\,\mathrm{k_B T}$. If this is the case the calculated score is smaller than the real one, and the binding strength of HS is underestimated. The algorithm checks whether this is the case and raises an error.

The last function is based on a physical interpretation, where we estimate the binding energy of the ligand and the protein. The approach is similar to our electrostatic interaction energy model in section 4.2.1. We assume that each point on the surface S represents a binding state of the ligand to the protein and compare these states with unbound states located in a volume without acting forces, i.e. $E_{unbound} = 0$. The difference in Gibbs free energy is then given as

$$\Delta G_{bind}(solv \rightarrow prot) = -RT \ln \left(\frac{1}{N_{unbound}} \sum_{j=1}^{N_S} \exp\left(-\frac{\Delta G_j(\vec{r})}{k_B T} \right) \right), \tag{4.22}$$

with the gas constant R, the number N_S of points on the surface and the number of points of the unbound reference state $N_{unbound}$, which can also be used to set the concentration

applied during the calculation. We set $N_{unbound}$ to yield concentrations of 1 mol/l of the protein and the ligand. The unit of the binding energy ΔG_{bind} is kJ/mol.

4.2.8 Alanine scanner

Our interaction energy model compares differences in concentration or, more precisely, differences in spatial probabilities of HS around proteins. We wanted to use our model to predict which residues are important for the binding of HSs to proteins. Alanine scans are a standard experimental procedure to assess the individual contribution of one amino acid [97]. Therefore, we replaced all occurrences of the two basic amino acids lysine and arginine, which are known to bind HS, with alanine using point mutations.
Due to their long side chains, the basic amino acids lysine and arginine are rather flexible. This flexibility can have large impacts on the electrostatic potential and the surface around the protein. For example, if the side chain is extended away from the protein it leads to a bump in the surface and the charge can unfold its full effect, whereas if the charge is close to the surface and close to the other side chains, its charge can be partially shielded. Therefore, we built ten structures for each point mutation with Modeller to account for different rotamer arrangements.

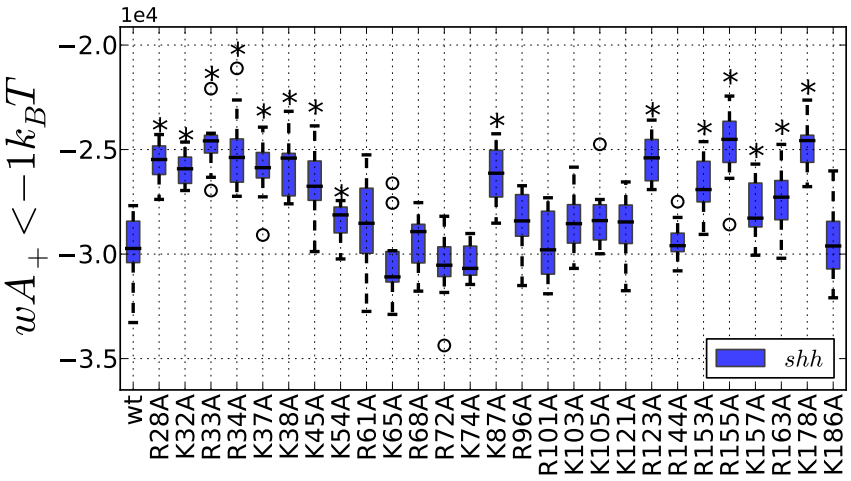


Figure 4.2: Alanine scan of Shh - The weighted surface scoring function wA_+ was used for the calculation. The asterisk denotes alanine mutations for which the distribution of the scoring function significantly differs from the wild type.

An example for an alanine scan is illustrated in figure 4.2. We used RPy2[1] to interface with the software R [131] and used the Wilcoxon Rank Sum and Signed Rank Test to

[1] http://rpy.sourceforge.net/rpy2.html

assess the significance of the difference of each point mutation compared with the wild type.

4.2.9 Verification of experimentally determined interacting regions on proteins

For the validation of the method we selected proteins from the PDB for which the binding residues to heparin/HS are known or which have been solved in complex with a heparin/HS fragment. In the following, a short description for each PDB entry is given. We used the visualization of isosurfaces at $-2\,k_BT$ to assess if the heparin/HS fragments or known binding residues are covered by the isosurfaces and thus that our predictions are correct. The results for all proteins are illustrated in Figure 4.3.

1fq9 GAG binding promotes the dimerization of fibroblast growth factor 2 complexed with fibroblast growth factor receptor 1 [120]. The structure has a chain break and missing residues in chain D. The missing residues were added using Modeller [119].

1gmo Binding of GAG to the hepatocyte growth factor protein enables biological activity and promotes dimer formation [76].

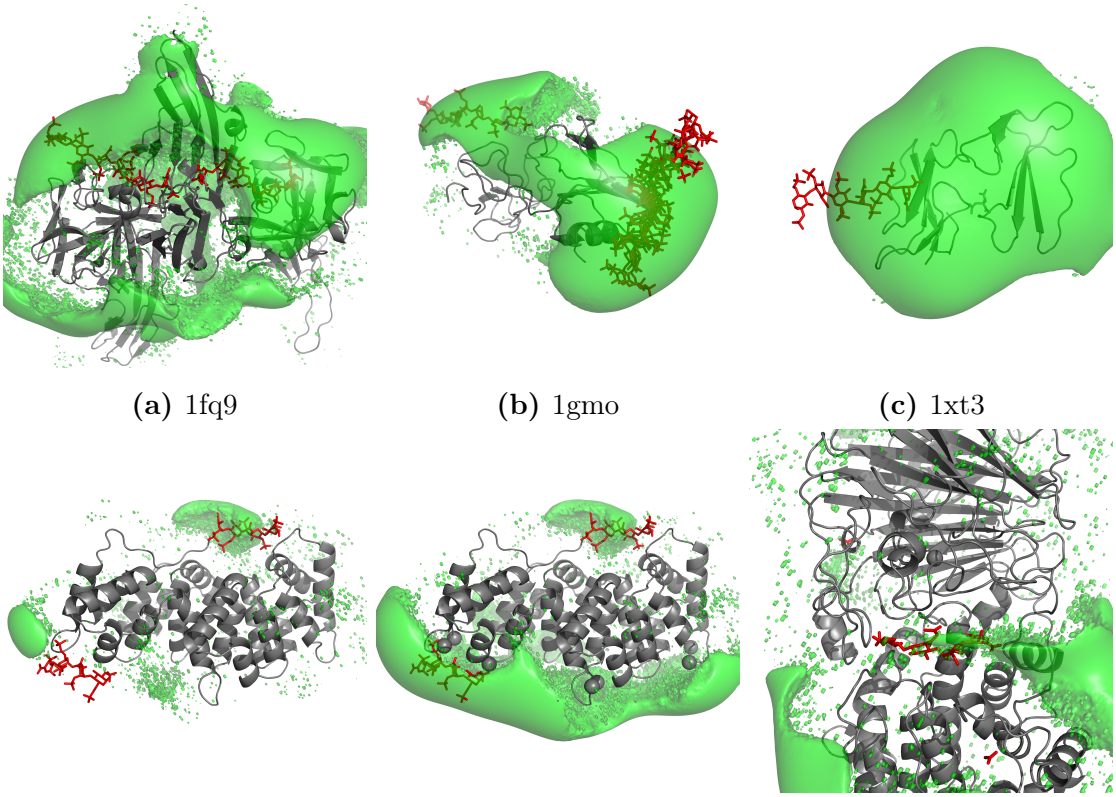
1xt3 Cytotoxin 3 is a cobra cardiotoxin (CTX). Binding of GAG does not only stabilize the protein, but also induces a conformational change, which leads to a citrate-mediated dimerization [73].

1g5n The membrane protein annexin V has been crystallized with heparin-derived tetrasaccharides, from which two distinct GAG binding sites could be identified. Moreover, annexin V also binds Ca^{2+} ions, which increase the GAG binding affinity in solution [10].

3e7j The enzyme heparinase II depolymerizes GAGs regardless of their sulfation pattern. Heparinase II has been complexed with an HS disaccharide product residing at the active/binding site [126].

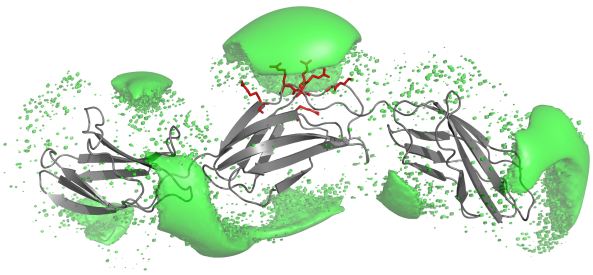
1fnh The protein fibronectin contains at least one GAG binding site. Evidence suggests the existence of a putative binding site, which is not part of this analysis. The crystal structure does not contain a GAG molecule. Therefore, residues of the known GAG binding site are shown as red sticks. [125]

We found agreement between the computational predictions of GAG binding regions around proteins and the experimental derived bindings sites of protein-GAG complexes by visualization of isosurfaces at an energy level of $-2\,k_BT$. Isosurfaces of favorable energies of all investigated structures were located over the experimentally derived GAG fragments/residues (see figure 4.3).
Our method was based on the premise that electrostatic forces are the major force driving the protein-HS interaction. From the comparison with experimental results, this assumption seems to be valid. The results suggest that our model is in general able to identify HS interaction sites around proteins.



(a) 1fq9 **(b)** 1gmo **(c)** 1xt3

(d) 1g5n (no calcium) **(e)** 1g5n (with calcium) **(f)** 3e7j



(g) 1fnh

Figure 4.3: Comparison of the predictions with experimental data - Energy isosurfaces at $-2\,k_BT$ around proteins for which the binding site of heparin/HS is known or which have been crystalized with heparin/HS. Either the binding site or the heparin/HS fragment is shown as red sticks.

4.3 Chemokines

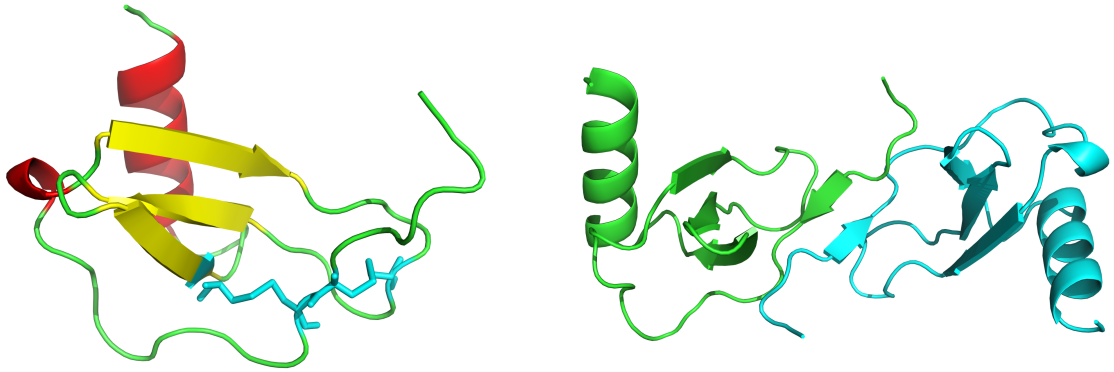
This section analyzes the interaction of chemokines with HS. Chemokines are **chemo** tactic cyto**kines**, meaning that they are signal proteins which have the ability to induce directed chemotaxis and thereby recruit leukocytes to the source of the inflammation/infection. The standard model is that chemokines bind to GAGs on the endothelial cell surface.

In humans, approximately 50 chemokines and 20 G-protein-coupled chemokine receptors have been identified which form a complex signaling network [141]. Experimental evidence suggests that only monomeric chemokines are able to activate receptors [112]. In solution, many chemokines can form dimers or oligomers without or upon binding to GAGs [44, 46, 54, 71]. The affinity of chemokines towards GAGs differs depending on the chemokine itself and the type of GAG and its composition [66, 129, 145]

The mechanism by which GAGs induce oligomerization in chemokines is still unknown. Predicting possible binding regions of GAGs around chemokines improves the understanding of the process. The presented analysis is performed for the chemokine CCL3 for which the oligomeric structure was recently solved by Ren et al. [114].

4.3.1 Structure of chemokines

Chemokines have a molecular weight ranging from 8 kDa to 12 kDa. Although chemokines are diverse in sequence they share a conserved tertiary structure. This fold is characterized by a disordered N-terminus of 6−10 amino acids, a long loop (N-loop) ending in an α-helix, followed by a three-stranded β-sheet, a C-terminal α-helix and disulfide bonds stabilizing the fold. All of our structures are based on PDB entry 2x69 [114]. The monomeric and dimeric structures of the chemokine representative CCL3 are shown in figures 4.4a and 4.4b.

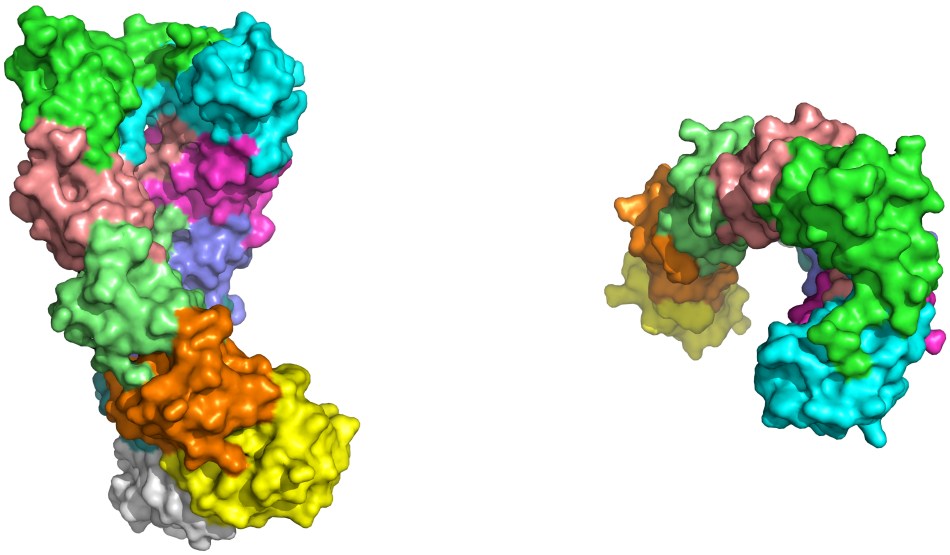


(a) CCL3 1mer **(b)** CCL3 2mer

Figure 4.4: Structure of CCL3 - Left: Cartoon representation of a CCL3 monomer. Cyan colored sticks illustrate the conserved cysteine bridges. Right: Cartoon representation of a CCL3 dimer. The two monomers are colored green and cyan.

CCL3 can reversibly form rod-shaped high molecular weight aggregates of 600 kDa [40, 114]. The multimeric structure of a CCL3 decamer is shown in figure 4.5. Ren et al. [114] found that adding heparin to CCL3 in solution reduced the average polymer size and increased their polydispersity while maintaining the rod shape. CCL4 is a homolog of CCL3 and has a sequence identity of 67 % and forms nearly identical rod shaped polymers [114]. The dimeric interaction of CCL4 and GAGs was previously analyzed by

Lortat-Jacob et al. [83]. Here, we extended the analysis to multimeric CCL3 to investigate the role of GAGs in multimeric chemokines.



(a) CCL3 10mer **(b)** CCL3 10mer

Figure 4.5: Multimeric structure of CCL3 - Surface representation of a CCL3 10mer. The view in the right image is rotated to highlight a channel, which is formed due to the multimerization on the surface of CCL3.

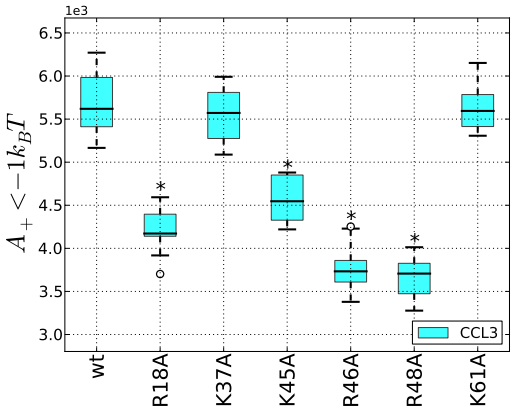
4.3.2 Alanine scan of CCL3

CCL3 was already studied by Koopmann and Krangel [63]. In subsection 4.2.7, we introduced five scoring functions for the characterization of the interaction of HS and proteins. In this subsection, we applied our implementation of an alanine scanner to CCL3 and tested which scoring function best resembled the experimental data by Koopmann and Krangel [63].

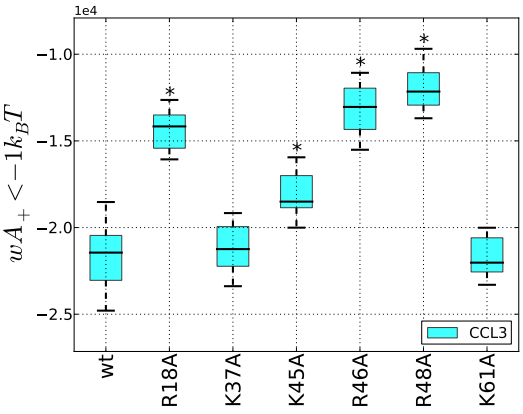
The results of the five scoring functions for dimeric CCL3 are illustrated in figure 4.6. Koopmann and Krangel [63] showed by site-directed mutagenesis that R18A, R46A and R48A failed to bind to heparin sepharose and that K45A had a weaker binding to heparin than the wild type. These experimental findings are best reproduced by the two scoring functions A_+ and wA_+. In both cases, the mutations are identified as significantly different from the wild type using a confidence interval of 95 % ($\alpha = 0.05$). The ranking according to the median of the distributions is also in agreement with the experiment. Mutation K45A is reported to elute at lower concentrations, which is reflected by its lower median compared to R18A, R46A and R48A.

In the case of the two volume-based heuristic functions (V_+ and wV_+) the four mutations are also separated from the wild type, but they are not the only significantly different mutations, and the ranking is also not in agreement with the experimental findings. The scoring function ΔG_{bind}, which uses a physical interpretation, fails to identify the experimental determined residues and is thus not suited for the characterization of the interaction. We reasoned that due to the absence of a cutoff, the energies below the cutoff introduced noise, which masked the signal.

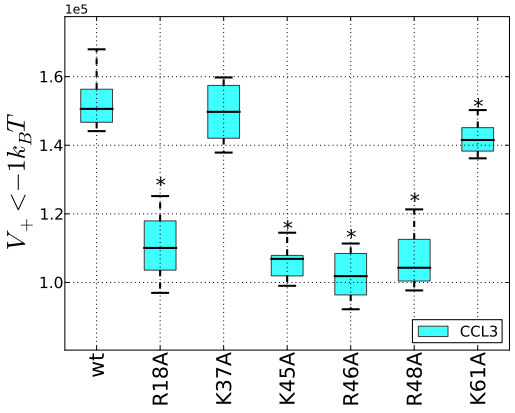
Although the performance of A_+ and wA_+ is almost identical, we chose wA_+ for the following analyses. We though that the incorporation of the energy values is advantageous especially for proteins with a high charge density, where the surface is close to the saturation of energy values $< -1\,\mathrm{k_B T}$.



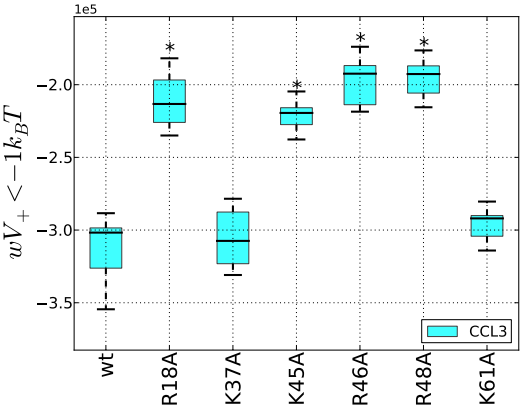
(a) A_+



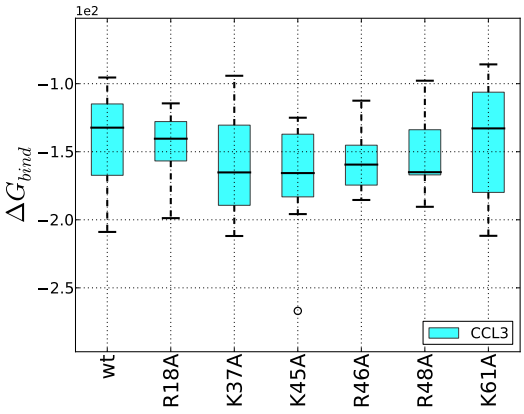
(b) wA_+



(c) V_+



(d) wV_+



(e) ΔG_{bind}

Figure 4.6: Alanine scan of CCL3 - Comparison of the five scoring functions applied to the alanine scan of dimeric CCL3.

4.3.3 How multimerization affects the binding of heparan sulfate to CCL3

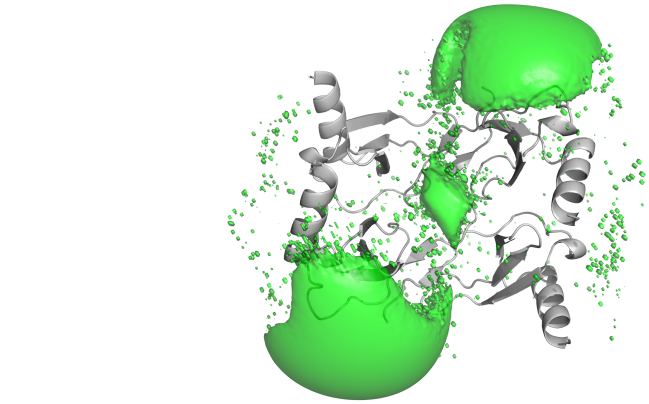
The previous subsection showed that the scoring function in combination with the underlying energy model is in good agreement with experimental studies. Here, we analyzed the energies around tetrameric CCL3 wild type and for all alanine mutations identified as significantly different from the wild type. For each mutant, we illustrated the isosurfaces at an energy level of $-2\,k_BT$ in figure 4.7. The volume enclosed by the isosurfaces is smaller for the mutants. Compared with the wild type the most obvious effect is the absence of favorable regions inside the channel of R46A and R48A. According to the experiment both mutants do not bind to HS [63]. Therefore, we hypothesize that the channel might be involved in the binding of HS to CCL3. In the case of R18A, our data cannot explain why this mutant does not bind to HS. Residue R18 is located at the interface of the dimers, hence the mutation might affect the binding capabilities of CCL3 in general. Unfortunately this was not tested by Koopmann and Krangel [63] because at the time of the study, the multimeric structure was not available and they reasoned R18 could not be part of a multimeric interface.

The interaction of HS and CCL3 was described by Hoogewerf et al. [46] and Ren et al. [114], though with different results. Both studies described that the addition of heparin had an impact on the degree of multimerization of CCL3. Hoogewerf et al. [46] found in their experiments that CCL3 was mostly dimeric and that the addition of low molecular weight heparin increased the molecular weight of CCL3. In contrast, Ren et al. [114] found in their experiments that CCL3 had an average polymer size ranging from 40-50 monomers and that the polymer size dropped to 20-30 units after adding heparin.

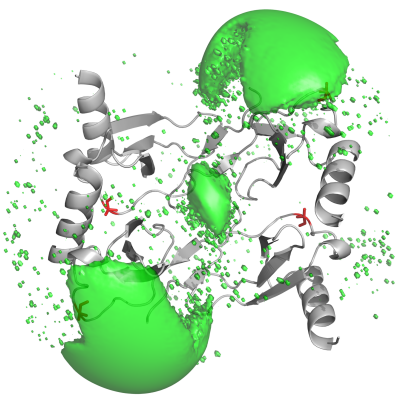
Unfortunately, there is no data available providing information about the type of heparin used in the experiments; especially the length of the heparin used would be of interest. Regarding the channel described above, we hypothesize that heparin serves as a string to which CCL3 dimers are attracted by electrostatic forces. Once they have attached to heparin they can form high-order multimers. The size of the multimers would be limited by the length of heparin or, in the case of HS, limited by the sulfation pattern.

The different results found by Ren et al. [114] and Hoogewerf et al. [46] could be explained with different initial conditions. Ren et al. [114] started with large polymers (ranging from 40 to 50), whereas in the experiment performed by Hoogewerf the wild type had a polymer

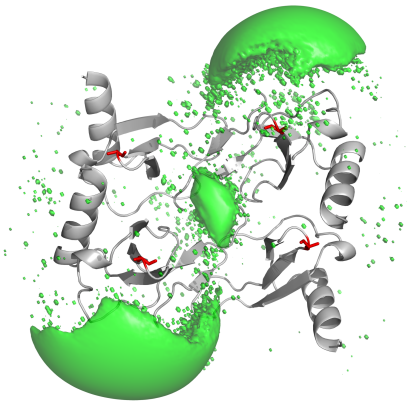
size ranging from one to three monomers. Assuming that the affinity of CCL3 towards HS is higher than the affinity towards itself, the experimental results would be explained with the given model.



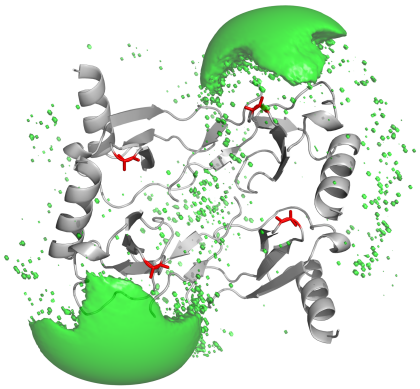
(a) wt



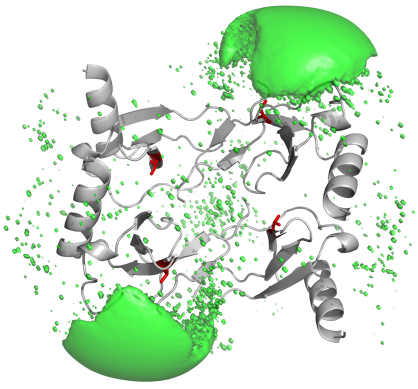
(b) R18A



(c) K45A



(d) R46A



(e) R48A

Figure 4.7: Energy isosurfaces of the CCL3 mutants - Energy isosurfaces at $-2\,k_BT$ for the CCL3 wild type and the four top ranked mutants. Mutants were ranked according to the mean of the distribution of the heuristic scoring function wA_+. The mutations are highlighted as red sticks.

4.4 Hedgehogs

Hedgehogs (Hhs) are morphogens, which control the growth and patterning in developing embryos. Hhs are emitted from a localized source to the extracellular space, where they form concentration gradients [146]. Prior to release to the extracellular space, a cholesterol is attached to the C-terminus and a palmitic acid resides at the N-terminus of Hh [106, 109]. After diffusion Hh binds to the membrane receptor Patched (Ptc), which results in the inhibition of Smoothend (Smo) and an activation of the Hh signal pathway, reviewed by Ryan and Chiang [118].

The mechanism of the signal transmission is still under discussion. Experiments performed by Bellaiche et al. [5] showed that diffusion and signaling of Hh was impaired in *Drosophila* if the enzyme Tout Velu (Ttv), which is associated with the biosynthesis of heparan sulfate proteoglycans (HSPGs) [132], was knocked out. In vertebrates, there exist three homologs of Hh named Sonic Hedgehog (Shh), Desert Hedgehog (Dhh) and Indian Hedgehog (Ihh). For Ihh it could be shown that a knock-down of the Exostosin gene, which is the mammalian homolog of Ttv, also influences the signaling range [65].

Furthermore, it was also suggested that HS mediates multimerization of Hh proteins [22, 135]. However, the multimerization process is still under discussion. One model of multimerization is based on the symmetry information from a crystal structure [99]. This model requires shedding of the N- and C-terminal lipid modifications, whereas Palm et al. [101] suggested that multimerization is based on the association of lipoproteins. Next to the release of lipoprotein associated Shh proteins, they also found that monomeric Shh without the lipid modifications was released from the cell.

It has already been established that Hhs contain a conserved Cardin-Weintraub (CW) motif [12] in the N-terminal loop, which binds to heparan sulfate [21, 41, 116] and is thought to mediate multimerization [30].

In this study, we analyzed the interactions of HS and Hhs. We compare the interactions across the Hh homologs and predict the residues involved in the binding of Hhs to HS.

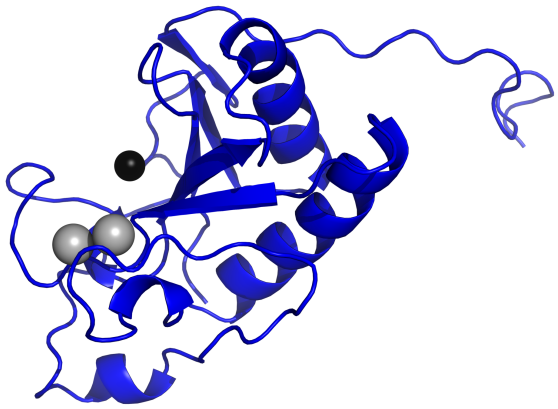


Figure 4.8: Structure of Shh - Cartoon representation of human Sonic Hedgehog. Zinc (black) and calcium (gray) ions are shown as spheres. The structure is based on PDB entry 3m1n [105] with calcium ions added from PDB entry 2wfx [7].

4.4.1 Structure of Hedgehogs

The structure of Hhs is illustrated for the human Hh homolog Sonic in figure 4.8. It consists of two α-helices, a mixed β-sheet of six strands connected by extensive loops and a small two stranded antiparallel β-sheet [43]. The N-terminal loop, which contains the CW motif, varies in length. Structurally, Hhs belong to a group of lysostaphin-type peptidase (LAS), which are metallopeptidases, although no enzymatic function has been reported for Shh so far. Nevertheless, Shh, Dhh and Ihh contain the same zinc center as the other LAS proteins. Only the Hh homolog found in fly has no zinc center and the zinc ligands are not conserved [43, 93]. In addition, all homologs have a second metal ion center, which can contain up to two calcium ions [94] and the ligands of the calcium ions are conserved in all homologs.

4.4.1.1 Structure preparation

There are different Hhs structures listed in the PDB, but there is no structure with a full-length N-terminal loop and both calcium ions. Therefore, we used the full length

crystal structure of PDB entry 3m1n [105] as a template and added the two calcium ions from PDB entry 2wfx [7] after structurally aligning it onto 3m1n. The structure is depicted in figure 4.8 and is referred to as $Shh_{ref,2ca}$ from now on.
Information about possible multimerization conformations of Shh are also based on PDB entry 3m1n [53, 105]. The PDB entry already contains a dimer of Shh illustrated in figure 4.9a. By applying the symmetry operations from the crystallographic experiment, another binding mode can be identified, which is shown in figure 4.9b. Combining these binding modes results in the formation of layer structures, which can again be added together, forming large oligomeric structures. An example for such a layer structure is shown in figure 4.9c.
Further, we used the software Modeller [119] for the modeling of all Hedgehog homologs. All models are based on $Shh_{ref,2ca}$. During the modeling process, we used the crystal symmetry as a constraint to keep the N-terminal loop fixed. Because the N-terminal loop has no geometrical constraints like an α-helix or a β-sheet, Modeller models the loop with a high flexibility leading to diverse conformations of the loop. Therefore, the change in the energy landscape due to the conformational flexibility of the tail can be in the same range as the change resulting from a new sequence modeled on the protein, and thus the change cannot be reliably related to the new sequence.

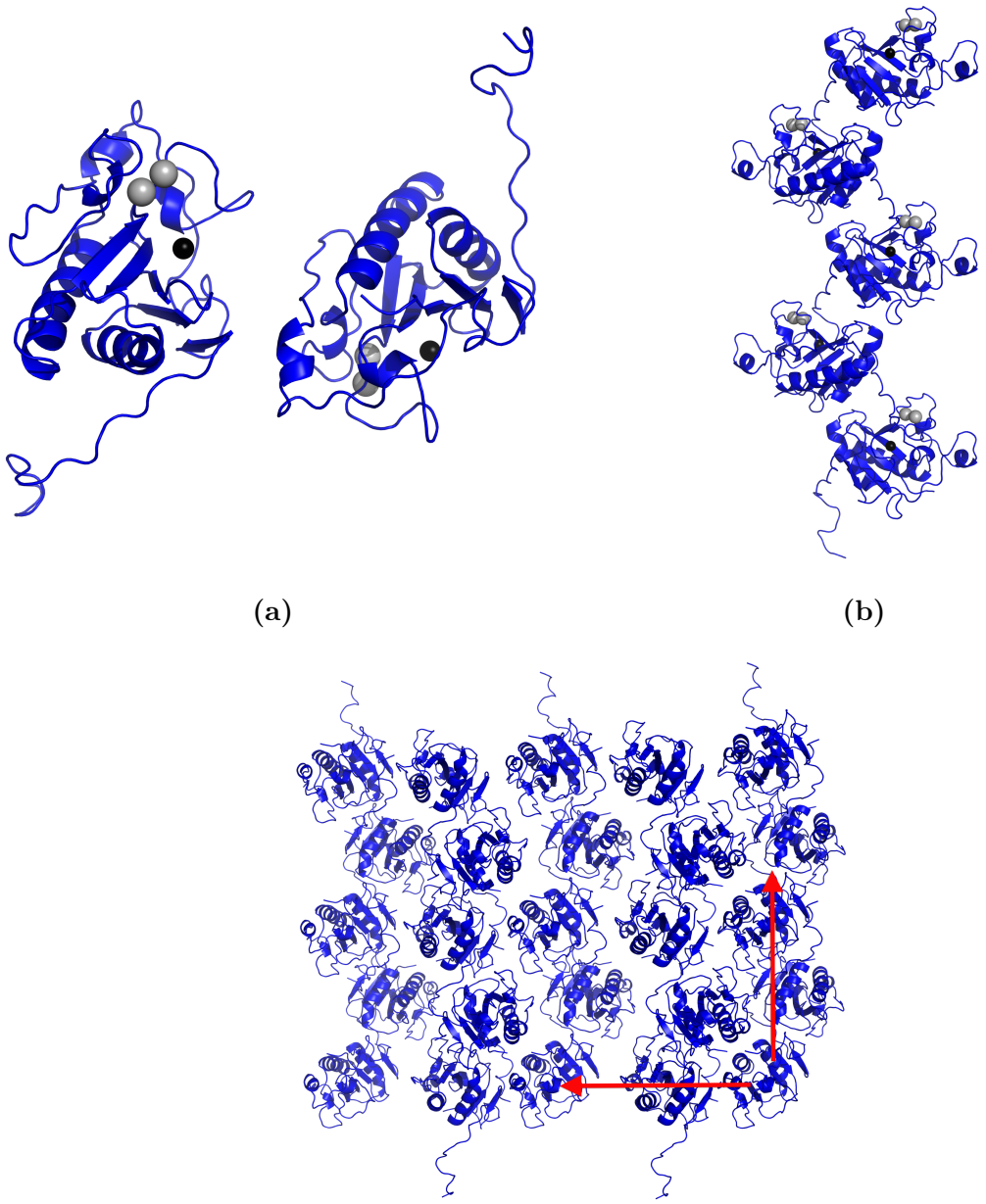


Figure 4.9: Multimerization modes of Hhs - (a) Shh dimer from the crystal structure 3m1n with two added calcium ions from the crystal structure 2wfx. (a) Shh pentamer according to another binding mode encoded in the symmetry operations of crystal structure 3m1n. (c) combination of the two binding modes depicted in figure (a) and (b) reveals a layered structure.

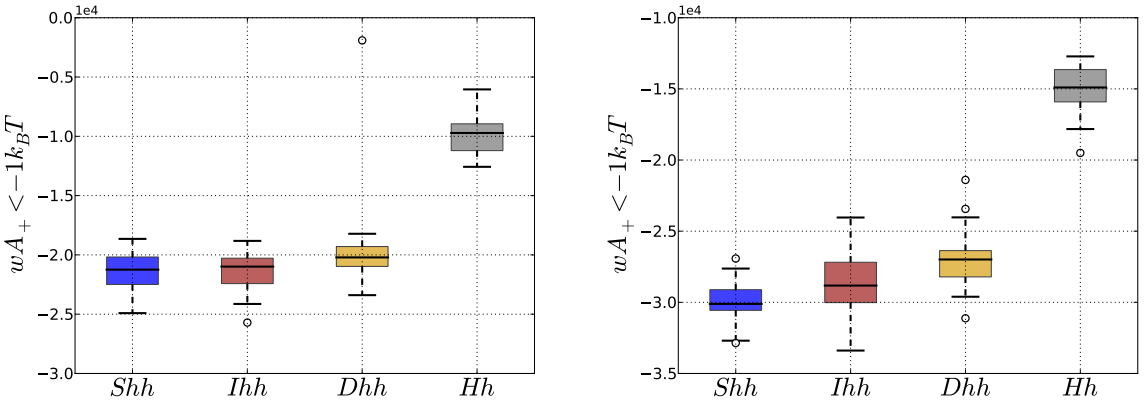
4.4.2 Analyzing the interaction between Hedgehog homologs and HS

Most of the experiments in the literature have focused on Shh and *Drosophila* Hedgehog. Here, we analyzed the binding of HS to all Hedgehog homologs. Therefore, we modeled all human Hedgehog homologs (Shh, Ihh and Dhh) and the *Drosophila* homolog (Hh) onto the reference structure $Shh_{ref,2ca}$. For each homolog, we constructed 21 models with and without calcium ions, which we used for the calculation of the difference in Gibbs free energy $\Delta G(s \rightarrow p)$. All models were aligned to the reference structure $Shh_{ref,2ca}$. The grid dimensions were set to $(193, 193, 193)$. We limited our analysis to the two states of Hhs defined by the absence or presence of the two calcium ions.

For the characterization of the HS interaction, we used the heuristic scoring function wA_+. The results for the four homologs are shown in figure 4.10. We used RPy2[1] to interface with the software R [131] and used the Wilcoxon Rank Sum and Signed Rank Test to access which homologs are significantly different from each other using a confidence interval of 95 % ($\alpha = 0.05$). If both calcium ions are present (figure 4.10b) all homologs are significantly different from each other. However, if no calcium ions are present Shh and Ihh are no longer different.

In their experiment Zhang et al. [151] compared the binding strength of HS to Shh and to Hh. They showed that HS binds to Shh, but that binding of HS to Hh was too weak to detect. The ranking of our results is in agreement with the experimental findings.

[1] http://rpy.sourceforge.net/rpy2.html



(a) **(b)**

Figure 4.10: Comparison of the four Hh homologs - Boxplots characterizing the binding of HS and the four Hh homologs in the absence of the two calcium ions (b) and in their presence (a).

4.4.3 Alanine scan of mammalian Hedgehog homologs

For the prediction of important residues, we used the heuristic scoring function wA_+. We performed the alanine scan for all three mammalian Hh homologs. We analyzed the *Drosophila* homolog in the following chapter. Mammalian Hhs harbor a conserved calcium binding site, which can bind up to two calcium ions, adding a charge of 4 e. We analyzed the structures in the absence (denoted by an index 0) and the presence (denoted by an index 2) of the two calcium ions.

For further validation of our predictions, we followed the idea of Lortat-Jacob et al. [83], who used sulfate ions from the crystallization buffer to validate their predictions of energetically favorable regions of sulfate groups around proteins.

Therefore, we retrieved all available Hh structures with sulfate ions (1vhh, 2wfq, 2wfr, 3k7g, 3k7h, 3k7i, 3k7j, 3m1n and 3mxw) from the PDB. Afterwards, we structurally aligned all structures to our reference model and extracted all sulfate groups. We compared the experimentally observed sulfate positions with our predictions.

The results of the alanine scan of the three homolog structures are shown in figure 4.11 along with a reference color bar[1]. All residues which differed significantly from the wild type are shown as sticks and are also listed in the appendix in table 5.10. The b-factor information in the PDB file has been used to assign a color to the residues depending on the deviation from the wild type. The formula for the coloring is given by

$$dev_i = \left| \frac{\overline{wV}_{+,i} - \overline{wV}_{+,wt}}{\overline{wV}_{+,max} - \overline{wV}_{+,wt}} \right|, \tag{4.23}$$

with the median of the wild type score $\overline{wV}_{+,wt}$, the median of the residue with the maximum deviation $\overline{wV}_{+,max}$ and the median of the residue $\overline{wV}_{+,i}$ for which the deviation dev_i from the wild type is calculated. The deviation ranges from 0 (no difference compared to the wild type) to 1 (maximum difference compared to the wild type).

For all mammalian homologs, we found that alanine mutations in the N-terminal loop are significantly different from the wild type and that these mutations are amongst the residues, for which we found the largest deviation to the wild type (colored as red sticks in figure 4.11). Furthermore, two sulfate groups are also located near these residues in the crystal structures. These results are in good agreement with the CW motif located in the N-terminal loop.

[1] http://www.pymolwiki.org/index.php/Palette_Colorbars

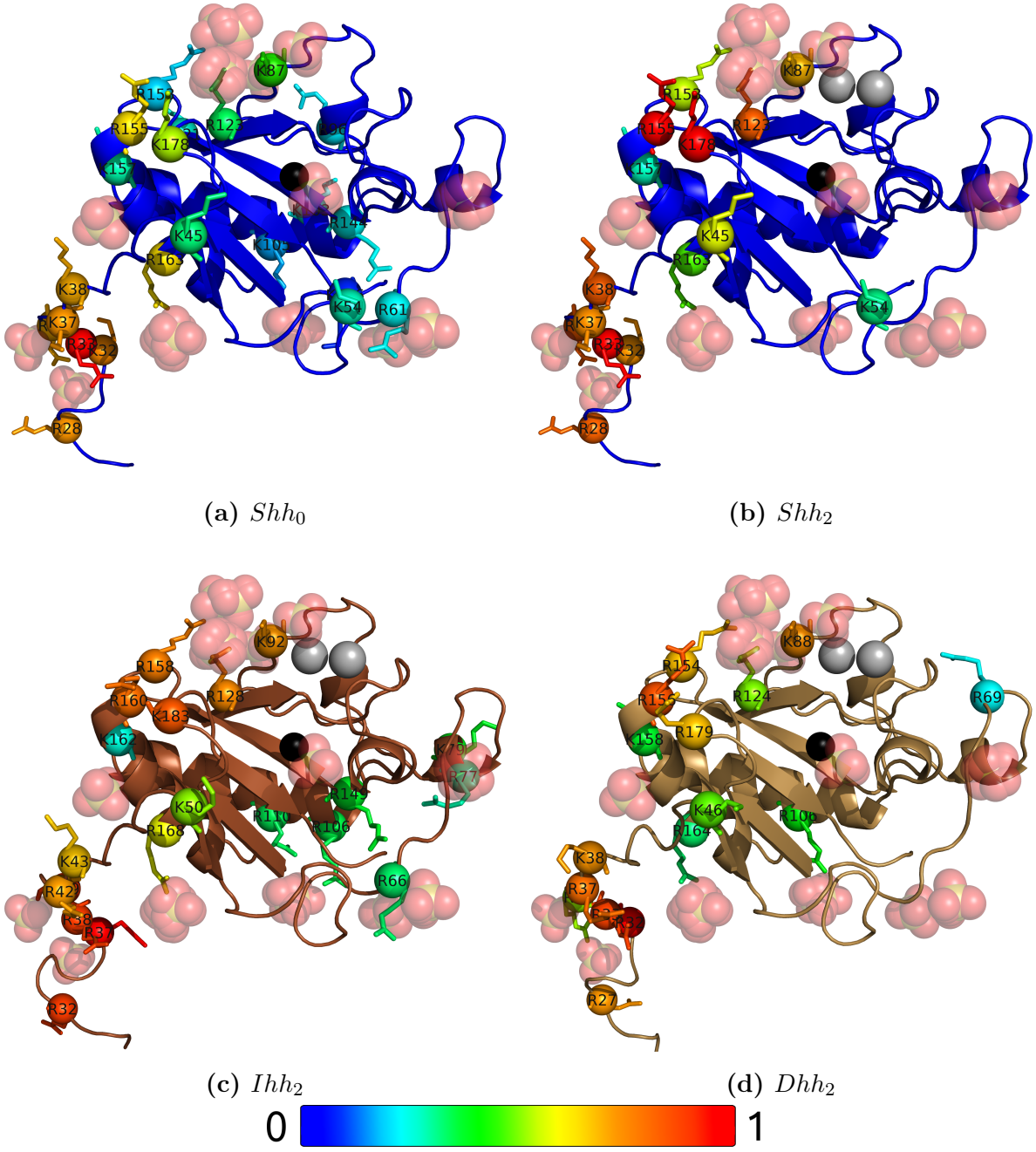


Figure 4.11: Alanine scan of the mammalian Hedgehog homologs - For Shh both states, with and without calcium ions, are illustrated. The other two homologs are only shown in the presence of calcium. Residues are shown as sticks and colored according to their deviation dev_i. For readability of the labels, the C_α atoms are shown as spheres. Additionally the sulfate groups extracted from all available Shh structures are shown as spheres.

Another interesting site of the homologs is located at the top of each subfigure in figure 4.11 and is made up of the residues K87, R123, R153, R155, K157 and K178 in Shh_2. This cluster is also found in the absence of calcium ions, although the deviations dev_i are lower. Again, several sulfate groups reside in close proximity to these residues. Therefore, we hypothesize that these residues constitute a putative binding site for HS in Hhs. Similar to the N-terminal CW motif, these residues might form a discontinuous CW motif. Residue K178 was already described as being crucial for HS binding by Chang et al. [14]. In their study, they also docked a undecasaccharide to Shh, which was placed across the CW motif and the putative CW motif. The distance between the CW motif and the discontinuous CW motif lies in the range from 3 nm to 5 nm. Zhang et al. [151] reported that at least an octasaccharide was required for binding to *Shh*. Our HS disaccharide has a length of around 1 nm, which would yield a length of around 4 nm for an octasaccharide. This length coincides with the distance between the two CW motifs. Therefore, we also hypothesize that HS might bind across both CW motifs. Ohlig et al. [99] reported that a partial removal of the CW motif impaired the binding of HS, suggesting that both binding sites are necessary for HS binding.

Another interesting residue is R163 (Shh), which is located between the CW motif located in the N-terminal loop and the putative CW motif. If HS binds to both motifs on a single protein, this residue is probably not part of the interaction, as it is too far away. However, if we align the sulfate groups to the multimeric structure based on the symmetry operations of 3m1n [105], R163 is in close proximity to the putative CW motif of the neighbor chain. Figure 4.12 shows that this residue is also in close proximity to several sulfate groups. We hypothesize that HS may occupy this space and enhance the binding of multimeric Shh.

For Shh the difference set of the two calcium states contains residues R61, R96, K103, K105, K121, R144 and K186, which are significantly different only in the absence of the two calcium ions, whereas the rest is identical. These residues have low deviations $dev < 0.26$ and are distributed over the protein, suggesting that they might be false positives. The putative CW motif exists regardless of the presence of calcium ions, although our model suggests that the presence of calcium ions enhances the binding of HS in this region. During the finalization of this work Whalen et al. [142] presented a complex of heparin bound to Shh in their study confirming the existence of the putative CW motif. They

found that all of the residues which we used for the definition of the putative CW motif are involved in binding of heparin to Shh.

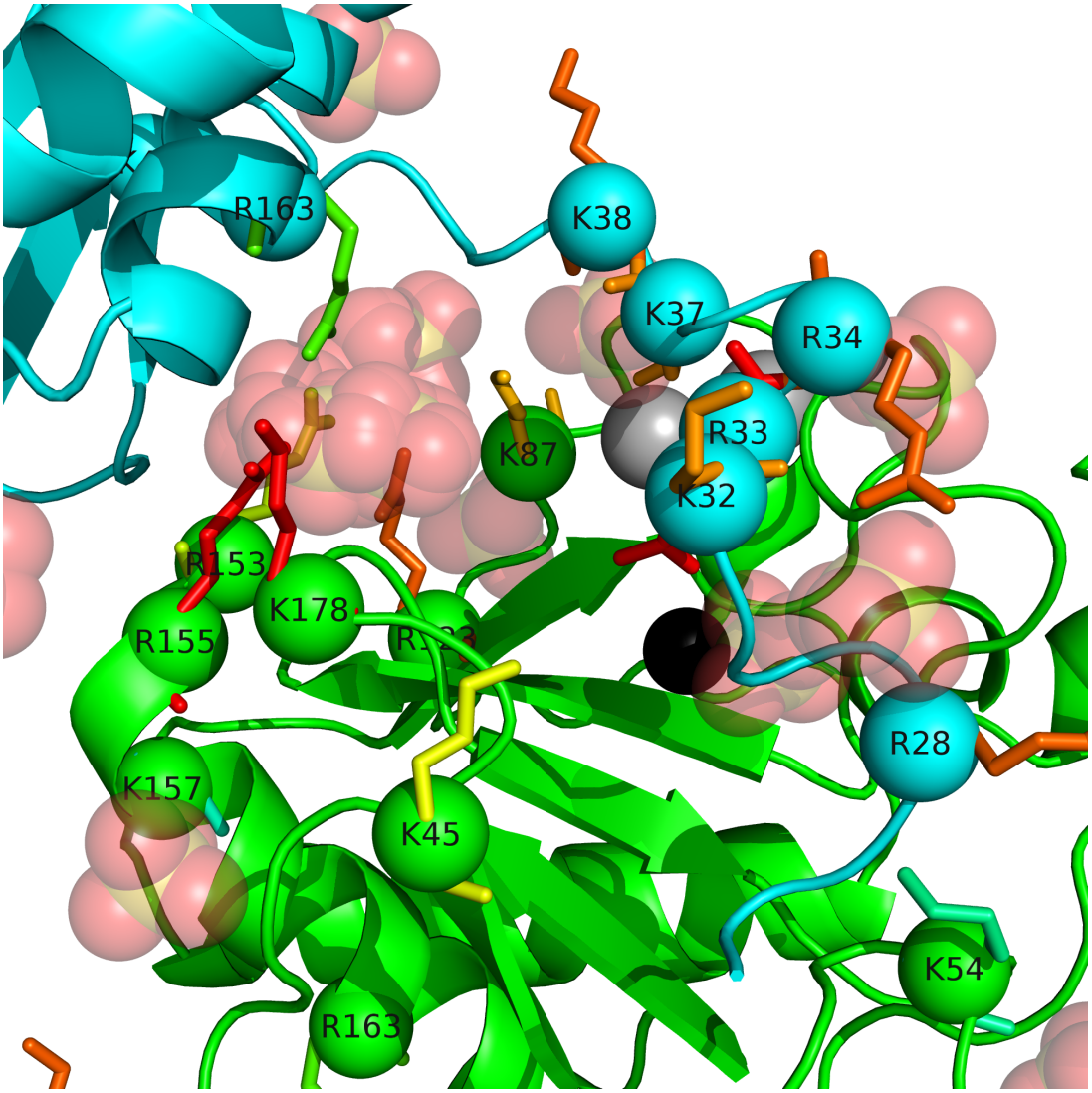
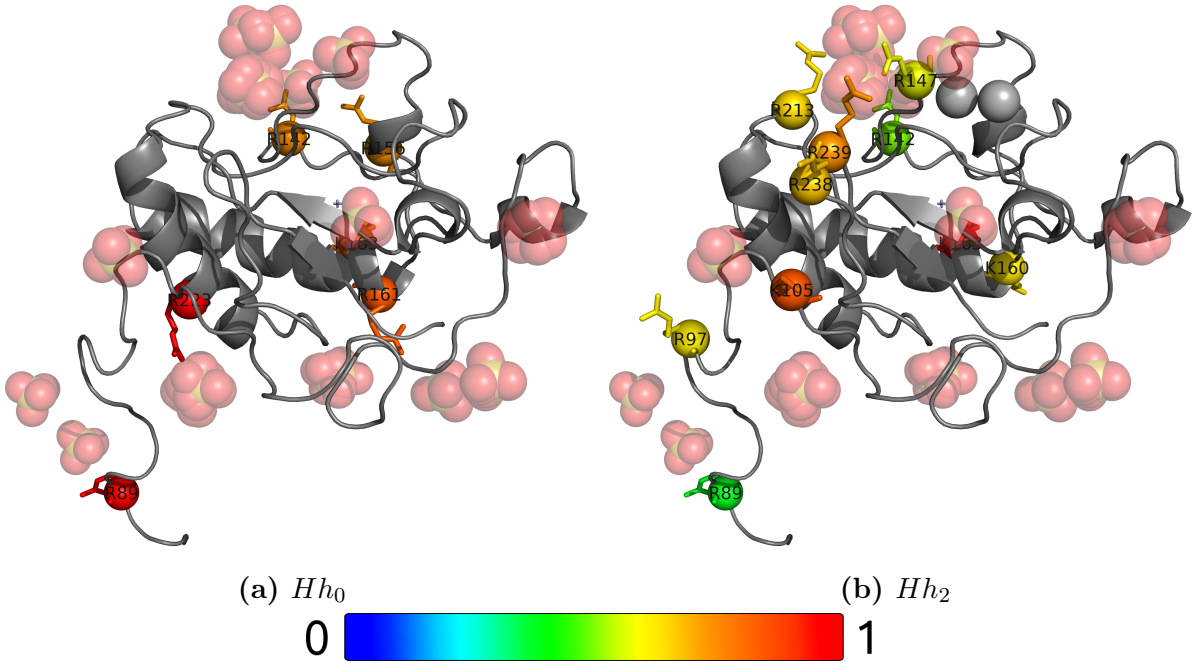


Figure 4.12: Multimer interactions - Second binding mode from the crystal structure 3m1n (compare figure 4.9b) in combination with the experimentally found sulfate groups and all residues, which have been identified with the alanine scanner. Residues are shown as sticks and colored according to their deviation dev_i. For readability of the labels, the C_α atoms are shown as spheres and colored according to their respective chains as green or cyan.

4.4.4 Alanine scan of *Drosophila* Hedgehog

In contrast to the mammalian Hh homologs, *Drosophila* Hedgehog does not have a zinc ion $Zn^{2+}$and thus its total charge is reduced by 2 e. We therefore chose to separate the analysis from the mammalian homologs. Here, we performed the alanine scan of *Drosophila* Hedgehog with and without calcium ions. The results are illustrated in figure 4.13, and all significantly different residues are listed in table 5.10.
McLellan et al. [93] found that in *Drosophila* residues K105, R147, R213 and R239 (Hh numbering) are involved in heparin binding. All of these residues are identified for Hh_2. For Hh_0, the number of residues identified as significantly different from the wild type is the fewest of all homologs and none of the ones found by McLellan et al. [93]. Hh_0 is also the protein with the lowest absolute charge of 2 e. In contrast, CCL3 contained −4 e and the lowest charge number for the mammalian Hhs is 4 e, because of the zinc ion Zn^{2+}. The heuristic scoring function wA_+ uses a threshold of $-1\,k_BT$ for the calculation. We reasoned that for proteins with a low number of charges the threshold is too high and thus the scoring function is not able to identify the important residues.



(a) Hh_0 **(b)** Hh_2

Figure 4.13: Effect of the calcium ions upon *Drosophila* Hedgehog - Comparison of the results of the alanine scanner in the absence (left) and presence (right) of the two calcium ions for Hh. Residues are shown as sticks and colored according to their deviation dev_i. For readability of the labels, the C_α atoms are shown as spheres. Additionally the experimentally derived sulfate groups are shown as spheres.

4.5 Summary and outlook

The electrostatic interaction energy model is capable of reproducing the experimental results reported in the experimental literature. Our model predicted an HS-binding region located in the N-terminal loop of all Hh homologs in agreement with other studies. Furthermore, a second cluster of predicted HS-binding residues could be identified. This cluster is discontinuous and located near the calcium binding site. The distance between the two clusters matches the length of an HS octasacharide, which was reported to be the minimal length required for binding of HS to Hhs [151], suggesting that both binding sites are involved in the binding process. These findings are conserved in all mammalian Hh homologs, and were recently confirmed for Shh by Whalen et al. [142]. We also analyzed the effect of the two calcium ions. Due to the spatial proximity of the putative HS-binding site, the calcium ions might increase the affinity to the putative CW motif.

Figure 4.13b shows that the deviation dev_i is a reliable descriptor for ranking the residues, i.e. residues with a high dev_i are in close proximity of the experimentally derived sulfate groups. The high-scoring residues cluster in the N-terminal loop and in the putative CW motif.

In the case of CCL3, we also found agreement of the predicted residues with the experimental literature. By using-high order multimers we could show that the effect of the alanine mutations is a reduction of favorable interaction regions in the channel of the multimer. Therefore, we hypothesize that the channel is involved in the binding to HS.

The current model is based only on electrostatic properties, which requires both, protein and ligand, to have an absolute total charge $\gg 0$. By incorporating further types of interactions, the model could be extended to other ligands. For the predictions, we used protein conformations only modeled with Modeller and GAG conformations created with PRODRG. One could use more elaborative methods for the sampling of conformations. For example, molecular dynamic (MD) simulations could be used to extract more realistic conformations of protein and the HS-ligand.

Another interesting application is the prediction of favorable regions around macromolecules. By incorporating a more sophisticated energy model, one could use different sulfation patterns of HS to screen macromolecules to understand the influence and importance of these patterns.

5

Appendix

Arf	Homeobox	RuBisCO_large	Tubulin
Cadherin	Hormone_recep	SH2	Y_phosphatase
Cytochrom_B_C	Kunitz_BPTI	SH3_1	fn3
Cytochrom_B_N	Lectin_C	Serpin	zf-C4
GTP_EFTU	RRM_1	Sushi	
Globin	Ras	Trypsin	

Table 5.1: Eukaryotic protein families.

Alpha_E1_glycop	Flavi_glycoprot	Pico_P1A	Rhv
Alpha_E2_glycop	HCV_capsid	Polyoma_coat	Rota_Capsid_VP6
Corona_S2	HN	RNase_H	VP4_haemagglut
Flavi_capsid	Late_protein_L1	RVP	
Flavi_glycop_C	Peptidase_C3	RVT_1	

Table 5.2: Viral protein families.

ABC_tran	GGDEF	Mur_ligase_C	Response_reg
ABM	GerE	Mur_ligase_M	RibD_C
AIRS	Globin	N6_Mtase	RimK
AIRS_C	Glycos_transf_1	N6_N4_Mtase	Rrf2
AP_endonuc_2	Glycos_transf_2	NMT1	SBP_bac_1
Amidohydro_3	Glyoxalase	NTP_transferase	SBP_bac_3
AraC_binding	GntR	Nitroreductase	SIS
ArsA_ATPase	HATPase_c	OEP	SLBB
AsnC_trans_reg	HD	OmpA	SLT
B12-binding	HTH_1	PAS	Sigma54_activat
BPD_transp_1	HTH_3	PASTA	Sigma70_r2
Bac_luciferase	HTH_5	PAS_3	Sigma70_r4
CMD	HTH_8	PD40	Sigma70_r4_2
COX1	HTH_IclR	PHP	Surf_Ag_VNR
CbiA	HisKA	PIN	T2SE
CheW	HlyD	PQQ	T2SF
CoA_transf_3	Hpt	PadR	TOBE
Cons_hypoth95	HxlR	ParBc	TOBE_2
Cytochrom_C	IclR	Pentapeptide	TP_methylase
DHH	IspD	Peptidase_M23	TctC
DHHA1	IstB_IS21	Peripla_BP_1	TetR_N
DNA_gyraseA_C	LacI	Peripla_BP_2	TonB
DegT_DnrJ_EryC1	LysR_substrate	Phage_integr_N	Toprim
EAL	MCPsignal	Phage_integrase	Trans_reg_C
FMN_red	MarR	PhoU	Transpeptidase
Fe-ADH	MerR	PilZ	Transposase_11
FecCD	MerR-DNA-bind	Plasmid_stabil	TrkA_N
Fer4	Methylase_S	Plug	TrmB
Fer4_NifH	MoCF_biosynth	ROK	UDPG_MGDP_dh_N
Flavin_Reduct	Molydop_binding	Radical_SAM	UTRA
Flavodoxin_2	Mur_ligase	Resolvase	YkuD

Table 5.3: Bacterial protein families.

re-weighting/x	0.5	0.6	0.7	0.8	0.9
DI_{PW}	150.6	154.3	153.3	152.5	151.2
DI_{MSA}	146.4	146.3	146.6	144.7	152.5

Table 5.4: TP-areas of the bacterial proteins.

re-weighting/x	0.5	0.6	0.7	0.8	0.9
DI_{PW}	162.0	164.9	167.1	167.2	166.2
DI_{MSA}	155.8	155.8	156.0	156.4	163.3

Table 5.5: TP-areas of the eukaryotic proteins.

re-weighting/x	0.5	0.6	0.7	0.8	0.9
DI_{PW}	72.4	79.0	79.5	78.6	79.1
DI_{MSA}	68.9	70.0	69.6	68.9	75.8

Table 5.6: TP-areas of the viral proteins.

i	i_{pdb}	j	j_{pdb}	DI
90	10	203	28	0.10842
111	16	206	29	0.05613
106	15	222	31	0.04317
207	30	225	34	0.04085
202	27	228	37	0.03590
118	20	198	23	0.03542
89	9	229	38	0.03300
119	21	198	23	0.02791
120	22	198	23	0.02620
203	28	229	38	0.02512
118	20	199	24	0.02121
90	10	223	32	0.02121
120	22	199	24	0.02121
89	9	227	36	0.02119
113	17	203	28	0.01977
117	19	201	26	0.01933
113	17	201	26	0.01908
203	28	227	36	0.01707
115	18	202	27	0.01386
118	20	200	25	0.01323

Table 5.7: Excerpt of the DI list of WD40 - Excerpt of the top 20 DI pairs of WD40 based on DI_{PW}. Pairs were sorted decreasingly with respect to their DI value and mapped to the sequence of the PDB structure 1yfq. Columns i and j refer to the columns of the alignment. Indices i_{pdb} and j_{pdb} are the positions of the alignment columns mapped to the sequence of PDB structure 1yfq.

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
1	283	OD	453	OD	0.37128
2	502	ID	607	GP41	0.24281
3	231	ID	267	OD	0.24210
4	747	GP41	758	GP41	0.22793
5	97	ID	275	OD	0.21283
6	13	SP	20	SP	0.21215
7	159	V2	174	V2	0.18100
8	360	OD	465	V5	0.17737
9	293	OD	337	OD	0.17296
10	92	ID	238	ID	0.16418
11	277	OD	352	OD	0.16132
12	308	V3	316	V3	0.14778
13	816	GP41	824	GP41	0.14248
14	49	ID	99	ID	0.14058
15	65	ID	208	ID	0.13247
16	825	GP41	833	GP41	0.13118
17	308	V3	315	V3	0.12682
18	211	ID	379	OD	0.12479
19	232	ID	268	OD	0.12257
20	231	ID	268	OD	0.12100
21	567	GP41	629	GP41	0.11996
22	219	ID	225	ID	0.11930
23	114	ID	202	BS	0.11632
24	275	OD	282	OD	0.11533
25	167	V2	192	V2	0.10877
26	230	ID	240	ID	0.10877
27	85	ID	229	ID	0.10662
28	290	OD	340	OD	0.10399
29	178	V2	195	V2	0.10118
30	106	ID	174	V2	0.09934
31	801	GP41	825	GP41	0.09326
32	133	V1	155	V1	0.09276
33	11	SP	21	SP	0.09273
34	325	V3	419	OD	0.09254
35	306	V3	321	V3	0.09185
36	425	BS	432	BS	0.09022
37	788	GP41	797	GP41	0.08965
38	770	GP41	783	GP41	0.08893

Continued on next page

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
39	182	V2	192	V2	0.08886
40	300	V3	442	OD	0.08818
41	290	OD	337	OD	0.08712
42	202	BS	432	BS	0.08705
43	557	GP41	567	GP41	0.08669
44	602	GP41	651	GP41	0.08571
45	845	GP41	851	GP41	0.08536
46	667	GP41	674	GP41	0.08519
47	269	OD	348	OD	0.08515
48	12	SP	21	SP	0.08388
49	287	OD	481	ID	0.08359
50	178	V2	194	V2	0.08202
51	192	V2	426	BS	0.08160
52	290	OD	344	OD	0.07801
53	805	GP41	853	GP41	0.07770
54	270	OD	277	OD	0.07685
55	46	ID	492	ID	0.07624
56	10	SP	21	SP	0.07395
57	333	OD	389	V4	0.07353
58	172	V2	305	V3	0.07272
59	335	OD	412	V4	0.07165
60	12	SP	30	SP	0.07133
61	740	GP41	796	GP41	0.07122
62	309	V3	317	V3	0.07048
63	360	OD	467	V5	0.06954
64	800	GP41	853	GP41	0.06918
65	154	V1	300	V3	0.06862
66	698	GP41	705	GP41	0.06799
67	121	BS	429	BS	0.06785
68	700	GP41	758	GP41	0.06784
69	12	SP	20	SP	0.06779
70	303	V3	323	V3	0.06772
71	632	GP41	640	GP41	0.06734
72	12	SP	23	SP	0.06702
73	500	ID	619	GP41	0.06635
74	677	GP41	683	GP41	0.06618
75	232	ID	269	OD	0.06551
76	273	OD	481	ID	0.06494
77	306	V3	316	V3	0.06477

Continued on next page

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
78	161	V2	172	V2	0.06418
79	293	OD	446	OD	0.06410
80	720	GP41	727	GP41	0.06381
81	456	OD	466	V5	0.06374
82	328	V3	334	OD	0.06368
83	726	GP41	736	GP41	0.06366
84	279	OD	474	OD	0.06331
85	721	GP41	732	GP41	0.06305
86	595	GP41	602	GP41	0.06304
87	761	GP41	769	GP41	0.06263
88	164	V2	170	V2	0.06259
89	353	OD	468	V5	0.06200
90	816	GP41	825	GP41	0.06108
91	175	V2	194	V2	0.06073
92	9	SP	21	SP	0.06026
93	518	GP41	553	GP41	0.06000
94	13	SP	19	SP	0.05991
95	725	GP41	731	GP41	0.05989
96	750	GP41	756	GP41	0.05901
97	305	V3	319	V3	0.05889
98	167	V2	177	V2	0.05814
99	10	SP	282	OD	0.05812
100	281	OD	365	OD	0.05786
101	309	V3	315	V3	0.05775
102	202	BS	315	V3	0.05774
103	232	ID	240	ID	0.05762
104	174	V2	429	BS	0.05762
105	723	GP41	731	GP41	0.05747
106	65	ID	379	OD	0.05727
107	10	SP	23	SP	0.05724
108	95	ID	236	ID	0.05708
109	784	GP41	800	GP41	0.05704
110	308	V3	317	V3	0.05702
111	565	GP41	646	GP41	0.05698
112	651	GP41	658	GP41	0.05660
113	346	OD	395	V4	0.05659
114	800	GP41	825	GP41	0.05651
115	295	OD	446	OD	0.05642
116	809	GP41	853	GP41	0.05630

Continued on next page

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
117	32	SP	500	ID	0.05625
118	471	V5	477	ID	0.05571
119	121	BS	202	BS	0.05516
120	619	GP41	646	GP41	0.05513
121	665	GP41	677	GP41	0.05455
122	11	SP	26	SP	0.05447
123	10	SP	20	SP	0.05406
124	7	SP	21	SP	0.05371
125	134	V1	154	V1	0.05346
126	25	SP	31	SP	0.05328
127	152	V1	181	V2	0.05321
128	720	GP41	750	GP41	0.05308
129	809	GP41	824	GP41	0.05280
130	289	OD	344	OD	0.05256
131	792	GP41	800	GP41	0.05230
132	106	ID	121	BS	0.05211
133	283	OD	471	V5	0.05168
134	158	V2	173	V2	0.05167
135	671	GP41	683	GP41	0.05131
136	269	OD	351	OD	0.05120
137	602	GP41	654	GP41	0.05097
138	300	V3	323	V3	0.05094
139	796	GP41	812	GP41	0.05060
140	624	GP41	632	GP41	0.05058
141	725	GP41	743	GP41	0.05044
142	114	ID	553	GP41	0.05032
143	792	GP41	798	GP41	0.05025
144	722	GP41	824	GP41	0.05018
145	195	V2	432	BS	0.05017
146	302	V3	323	V3	0.05016
147	133	V1	152	V1	0.05008
148	700	GP41	746	GP41	0.05001
149	46	ID	632	GP41	0.04976
150	788	GP41	805	GP41	0.04973
151	162	V2	195	V2	0.04952
152	25	SP	722	GP41	0.04950
153	24	SP	722	GP41	0.04931
154	23	SP	29	SP	0.04916
155	301	V3	323	V3	0.04897

Continued on next page

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
156	164	V2	195	V2	0.04895
157	62	ID	209	ID	0.04860
158	10	SP	809	GP41	0.04848
159	746	GP41	758	GP41	0.04839
160	788	GP41	800	GP41	0.04838
161	20	SP	26	SP	0.04832
162	801	GP41	824	GP41	0.04825
163	7	SP	20	SP	0.04824
164	183	V2	194	V2	0.04820
165	167	V2	309	V3	0.04819
166	121	BS	315	V3	0.04816
167	244	ID	629	GP41	0.04815
168	753	GP41	762	GP41	0.04811
169	369	OD	429	BS	0.04807
170	167	V2	426	BS	0.04806
171	458	OD	466	V5	0.04781
172	22	SP	29	SP	0.04779
173	749	GP41	758	GP41	0.04776
174	288	OD	341	OD	0.04758
175	270	OD	341	OD	0.04752
176	295	OD	413	V4	0.04736
177	236	ID	792	GP41	0.04734
178	793	GP41	804	GP41	0.04734
179	746	GP41	756	GP41	0.04729
180	720	GP41	796	GP41	0.04729
181	236	ID	275	OD	0.04719
182	499	ID	605	GP41	0.04718
183	492	ID	612	GP41	0.04713
184	588	GP41	646	GP41	0.04705
185	23	SP	853	GP41	0.04704
186	9	SP	22	SP	0.04701
187	307	V3	319	V3	0.04694
188	177	V2	192	V2	0.04628
189	232	ID	271	OD	0.04611
190	379	OD	443	OD	0.04582
191	722	GP41	746	GP41	0.04580
192	174	V2	333	OD	0.04550
193	92	ID	633	GP41	0.04547
194	256	OD	265	OD	0.04521

Continued on next page

n	i_{HXB2}	i_{domain}	j_{HXB2}	j_{domain}	DI
195	232	ID	238	ID	0.04517
196	845	GP41	854	GP41	0.04515
197	275	OD	474	OD	0.04512
198	281	OD	353	OD	0.04508
199	750	GP41	758	GP41	0.04497
200	295	OD	444	OD	0.04465

Table 5.9: Top 200 predicted DI pairs of HIV-1 Env - The table contains all 200 predicted DI pairs predicted with DI_{PW} using a re-weighting threshold $x = 0.8$.

02A1	BF	BG	A2D	BC	47_BF	01F2	AGU
6	29	3	1	40	2	1	1
02D	02G	31_BC	02A	02C	02B	02O	22_01A1
3	2	2	3	1	1	1	7
21_A2D	02U	07_BC	02_AG	13_cpx	A1U	D	H
3	3	77	122	5	5	116	3
F2KU	DF1G	04_cpx	GJ	01AF2U	27_cpx	01B	01C
1	1	8	1	1	3	35	2
A1A2D	06A1	36_cpx	02AG	09A	16_A2D	11_cpx	17_BF
3	2	2	1	1	2	11	5
BCF1	43_02G	01DU	25_cpx	45_cpx	06_cpx	0206	CF1
1	1	1	2	3	11	4	2
15_01B	42_BF	0107	AF2	0209	C	49_cpx	G
4	2	1	1	4	1339	5	59
01A1G	K	GKU	O	33_01B	A1GHU	34_01B	08_BC
1	2	2	27	6	1	2	37
54_01B	02GK	A1A2CD	0213	28_BF	14_BG	U	39_BF
1	1	1	1	5	5	15	3
A1CD	ACD	A1B	20_BG	AKU	51_01B	37_cpx	A2C
6	3	3	2	1	1	3	3
01ADF2	A2G	12_BF	38_BF1	01BC	JKU	BF1	02BG
1	1	7	4	2	1	34	1
35_AD	AF2G	40_BF	B	A1F1	A1F2	F	CG
15	1	4	1659	1	1	1	1
CD	02A1U	CU	F1	F2	0102A	03_AB	23_BG
21	4	1	33	8	1	4	1
05_DF	AGKU	CF1U	DF	DG	01A1	BFG	48_01B
3	1	1	1	2	3	1	3
DU	01_AE	46_BF	10_CD	09_cpx	-	18_cpx	A1
1	345	7	3	5	175	1	200
A2	A1DK	29_BF	32_06A1	A	AC	AD	AG
4	1	7	1	49	17	17	1
0225	A1D	A1G	AU	A1C	24_BG	N	A1H
1	41	7	3	30	2	2	1
AHJU	A1GU	26_AU	A1CDGKU	0708	07B	A1GJ	
2	1	4	1	3	1	1	

Table 5.8: HIV subtypes - List of all subtypes included in our dataset and their corresponding frequencies.

Shh_0	R28A K32A R33A R34A K37A K38A K45A K54A R61A K87A R96A K103A K105A K121A R123A R144A R153A R155A K157A R163A K178A K186A
Shh_2	R28A K32A R33A R34A K37A K38A K45A K54A K87A R123A R153A R155A K157A R163A K178A
Ihh_0	R32A R37A R38A R39A R42A K43A K50A R66A R106A R110A K126A R128A R149A R158A R160A K162A R168A K183A
Ihh_2	R32A R37A R38A R39A R42A K43A K50A R66A R77A K79A K92A R106A R110A R128A R149A R158A R160A K162A R168A K183A
Dhh_0	R27A R32A R33A R34A R37A K38A R69A R73A K158A R164A R179A
Dhh_2	R27A R32A R33A R34A R37A K38A K46A R69A K88A R106A R124A R154A R156A K158A R164A R179A
Hh_0	R89A R142A R156A R161A K163A R223A
Hh_2	R89A R97A K105A R142A R147A K160A K163A R213A R238A R239A

Table 5.10: Significant residues of all Hh homologs - Overview over all alanine mutations for which the distribution of the scoring function differs significantly from the wild type.

References

[1] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. *Electrostatics of nanosystems: application to microtubules and the ribosome. Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037–41, 2001. ISSN 0027-8424. doi:10.1073/pnas.181342398. 57

[2] Battiste, J. L. and Wagner, G. *Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. Biochemistry*, 39(18):5355–65, 2000. ISSN 0006-2960. doi:10.1021/bi000060h. 6

[3] Bayer, E., Goettsch, S., Mueller, J. W., Griewel, B., Guiberman, E., Mayr, L. M., and Bayer, P. *Structural analysis of the mitotic regulator hPin1 in solution: insights into domain architecture and substrate binding. The Journal of biological chemistry*, 278(28):26183–93, 2003. ISSN 0021-9258. doi:10.1074/jbc.M300721200. 4

[4] Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. *Cython: The Best of Both Worlds. Computing in Science & Engineering*, 13(2):31–39, 2011. ISSN 1521-9615. doi:10.1109/MCSE.2010.118. 25, 56

[5] Bellaiche, Y., The, I., and Perrimon, N. *Tout-velu is a Drosophila homologue of the putative tumour suppressor EXT-1 and is needed for Hh diffusion. Nature*, 394(6688):85–8, 1998. ISSN 0028-0836. doi:10.1038/27932. 74

[6] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. *The Protein Data Bank. Nucleic acids research*, 28(1):235–42, 2000. ISSN 0305-1048. 4

[7] Bishop, B., Aricescu, a. R., Harlos, K., O'Callaghan, C. a., Jones, E. Y., and Siebold, C. *Structural insights into hedgehog ligand sequestration by the human hedgehog-interacting*

protein HHIP. Nature structural & molecular biology, 16(7):698–703, 2009. ISSN 1545-9985. doi:10.1038/nsmb.1607. 75, 76

[8] Brandts, J. F., Halvorson, H. R., and Brennan, M. *Consideration of the Possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. Biochemistry*, 14(22):4953–63, 1975. ISSN 0006-2960. 4

[9] Burger, L. and van Nimwegen, E. *Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS computational biology*, 6(1):e1000633, 2010. ISSN 1553-7358. doi:10.1371/journal.pcbi.1000633. 18

[10] Capila, I., Hernáiz, M. J., Mo, Y. D., Mealy, T. R., Campos, B., Dedman, J. R., Linhardt, R. J., and Seaton, B. a. *Annexin V–heparin oligosaccharide complex suggests heparan sulfate–mediated assembly on cell surfaces. Structure (London, England : 1993)*, 9(1):57–64, 2001. ISSN 0969-2126. 63

[11] Capila, I. and Linhardt, R. R. J. *Heparin - Protein Interactions. Angewandte Chemie International ...*, 41(3):391–412, 2002. ISSN 1433-7851. 49

[12] Cardin, A. D. and Weintraub, H. J. *Molecular modeling of protein-glycosaminoglycan interactions. Arteriosclerosis (Dallas, Tex.)*, 9(1):21–32, 1989. ISSN 0276-5047. doi:10.1161/01.ATV.9.1.21. 74

[13] Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. *Core structure of gp41 from the HIV envelope glycoprotein. Cell*, 89(2):263–73, 1997. ISSN 0092-8674. 38

[14] Chang, S.-C., Mulloy, B., Magee, A. I., and Couchman, J. R. *Two distinct sites in sonic Hedgehog combine for heparan sulfate interactions and cell signaling functions. The Journal of biological chemistry*, 286(52):44391–402, 2011. ISSN 1083-351X. doi:10.1074/jbc.M111.285361. 82

[15] Chen, H. I. and Sudol, M. *The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. Proceedings of the National Academy of Sciences of the United States of America*, 92(17):7819–23, 1995. ISSN 0027-8424. 3

[16] Cheng, J., Li, J., Wang, Z., Eickholt, J., and Deng, X. *The MULTICOM toolbox for protein structure prediction. BMC bioinformatics*, 13:65, 2012. ISSN 1471-2105. doi:10.1186/1471-2105-13-65. 1

[17] Chuang, W.-l., Christ, M. D., and Rabenstein, D. L. *Determination of the Primary Structures of Heparin- and Heparan Sulfate-Derived Oligosaccharides Using Band-Selective Homonuclear-Decoupled Two-Dimensional 1 H NMR Experiments. Analytical Chemistry*, 73(10):2310–2316, 2001. ISSN 0003-2700. doi:10.1021/ac0100291. 49

[18] Cohen, J. *Structural biology. Is high-tech view of HIV too good to be true? Science (New York, N.Y.)*, 341(6145):443–4, 2013. ISSN 1095-9203. doi:10.1126/science.341.6145.443. 38

[19] Combet, J., Rawiso, M., Rochas, C., Hoffmann, S., and BoueÌĄ, F. *Structure of Polyelectrolytes with Mixed Monovalent and Divalent Counterions: SAXS Measurements and PoissonâĹŠBoltzmann Analysis. Macromolecules*, 44(8):3039–3052, 2011. ISSN 0024-9297. doi:10.1021/ma102226v. 53

[20] Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, 2001. 10

[21] Desbordes, S. C. and Sanson, B. *The glypican Dally-like is required for Hedgehog signalling in the embryonic epidermis of Drosophila. Development (Cambridge, England)*, 130(25):6245–55, 2003. ISSN 0950-1991. doi:10.1242/dev.00874. 74

[22] Dierker, T., Dreier, R., Migone, M., Hamer, S., and Grobe, K. *Heparan sulfate and transglutaminase activity are required for the formation of covalently cross-linked hedgehog oligomers. The Journal of biological chemistry*, 284(47):32562–71, 2009. ISSN 1083-351X. doi:10.1074/jbc.M109.044867. 74

[23] Dobzhansky, T. *Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudoobscura. Genetics*, 35(3):288–302, 1950. ISSN 0016-6731. 18

[24] Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., and Baker, N. a. *PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic acids research*, 35(Web Server issue):W522–5, 2007. ISSN 1362-4962. doi:10.1093/nar/gkm276. 57

[25] Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society*, 125(7):1731–7, 2003. ISSN 0002-7863. doi:10.1021/ja026939x. 1

[26] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge Univ Pr, 1998. 22

[27] Dybowski, J. N. *Development of a method for optimal superposition of pairs of similar macromolecules.* diploma thesis, Fachhochschule Bingen, 2006. 56

[28] Eddy, S. R. *Profile hidden Markov models. Bioinformatics (Oxford, England)*, 14(9):755–63, 1998. ISSN 1367-4803. 25

[29] Esko, J., Kimata, K., and Lindahl, U. *Essentials of Glycobiology.* Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, New York, 2nd edition, 2009. ISBN 10:0-87969-559-5. 48, 49

[30] Farshi, P., Ohlig, S., Pickhinke, U., Höing, S., Jochmann, K., Lawrence, R., Dreier, R., Dierker, T., and Grobe, K. *Dual roles of the Cardin-Weintraub motif in multimeric Sonic hedgehog. The Journal of biological chemistry*, 286(26):23608–19, 2011. ISSN 1083-351X. doi:10.1074/jbc.M110.206474. 74

[31] Fischer, G. and Aumüller, T. *Regulation of peptide bond cis/trans isomerization by enzyme catalysis and its implication in physiological processes. Reviews of physiology, biochemistry and pharmacology*, 148:105–50, 2003. ISSN 0303-4240. doi:10.1007/s10254-003-0011-3. 3

[32] Fischer, S., Dunbrack, R. L., and Karplus, M. *Cis-Trans Imide Isomerization of the Proline Dipeptide. Journal of the American Chemical Society*, 116(26):11931–11937, 1994. ISSN 0002-7863. doi:10.1021/ja00105a036. 3

[33] Fitch, W. M. and Markowitz, E. *An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochemical genetics*, 4(5):579–93, 1970. ISSN 0006-2928. 18

[34] Frigo, M. and Johnson, S. *The Design and Implementation of FFTW3. Proceedings of the IEEE*, 93(2):216–231, 2005. ISSN 0018-9219. doi:10.1109/JPROC.2004.840301. 56

[35] Gabb, H. a., Jackson, R. M., and Sternberg, M. J. *Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of molecular biology*, 272(1):106–20, 1997. ISSN 0022-2836. doi:10.1006/jmbi.1997.1203. 50

[36] Gabdoulline, R. R. and Wade, R. C. *Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. Journal of molecular biology*, 306(5):1139–55, 2001. ISSN 0022-2836. doi:10.1006/jmbi.2000.4404. 50

[37] Gallagher, J. T. *Heparan sulfate: growth control with a restricted sequence menu. Journal of Clinical Investigation*, 108(3):357–361, 2001. ISSN 0021-9738. doi:10.1172/JCI200113713. 49, 58

[38] Gillespie, J. R. and Shortle, D. *Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. Journal of molecular biology*, 268(1):158–69, 1997. ISSN 0022-2836. doi:10.1006/jmbi.1997.0954. 13

[39] Göbel, U., Sander, C., Schneider, R., and Valencia, A. *Correlated mutations and residue contacts in proteins. Proteins*, 18(4):309–17, 1994. ISSN 0887-3585. doi:10.1002/prot.340180402. 18

[40] Graham, G. J., MacKenzie, J., Lowe, S., Tsang, M. L., Weatherbee, J. A., Issacson, A., Medicherla, J., Fang, F., Wilkinson, P. C., and Pragnell, I. B. *Aggregation of the chemokine MIP-1 alpha is a dynamic and reversible phenomenon. Biochemical and biological analyses. The Journal of biological chemistry*, 269(7):4974–8, 1994. ISSN 0021-9258. 67

[41] Grobe, K., Inatani, M., Pallerla, S. R., Castagnola, J., Yamaguchi, Y., and Esko, J. D. *Cerebral hypoplasia and craniofacial defects in mice lacking heparan sulfate Ndst1 gene function. Development (Cambridge, England)*, 132(16):3777–86, 2005. ISSN 0950-1991. doi:10.1242/dev.01935. 74

[42] Guttman, M., Kahn, M., Garcia, N. K., Hu, S.-L., and Lee, K. K. *Solution structure, conformational dynamics, and CD4-induced activation in full-length, glycosylated, monomeric HIV gp120. Journal of virology*, 86(16):8750–64, 2012. ISSN 1098-5514. doi:10.1128/JVI.07224-11. 38, 41

[43] Hall, T. M., Porter, J. A., Beachy, P. A., and Leahy, D. J. *A potential catalytic site revealed by the 1.7-A crystal structure of the amino-terminal signalling domain of Sonic hedgehog. Nature*, 378(6553):212–6, 1995. ISSN 0028-0836. doi:10.1038/378212a0. 75

[44] Handel, T. M., Johnson, Z., Crown, S. E., Lau, E. K., and Proudfoot, a. E. *Regulation of protein function by glycosaminoglycans–as exemplified by chemokines. Annual review of biochemistry*, 74:385–410, 2005. ISSN 0066-4154. doi:10.1146/annurev.biochem.72.121801.161747. 49, 66

[45] Harris, R. C., Boschitsch, A. H., and Fenley, M. O. *Influence of Grid Spacing in Poisson-Boltzmann Equation Binding Energy Estimation. Journal of Chemical Theory*

and Computation, page 130614130709005, 2013. ISSN 1549-9618. doi:10.1021/ct300765w. 57

[46] Hoogewerf, a. J., Kuschert, G. S., Proudfoot, a. E., Borlat, F., Clark-Lewis, I., Power, C. a., and Wells, T. N. *Glycosaminoglycans mediate cell surface oligomerization of chemokines. Biochemistry*, 36(44):13570–8, 1997. ISSN 0006-2960. doi:10.1021/bi971125s. 66, 71

[47] Hopf, T. a., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. *Three-dimensional structures of membrane proteins from genomic sequencing. Cell*, 149(7):1607–21, 2012. ISSN 1097-4172. doi:10.1016/j.cell.2012.04.012. 26

[48] Hu, G., Liu, J., Taylor, K. a., and Roux, K. H. *Structural comparison of HIV-1 envelope spikes with and without the V1/V2 loop. Journal of virology*, 85(6):2741–50, 2011. ISSN 1098-5514. doi:10.1128/JVI.01612-10. 38

[49] Huang, C.-C., Lam, S. N., Acharya, P., Tang, M., Xiang, S.-H., Hussan, S. S.-U., Stanfield, R. L., Robinson, J., Sodroski, J., Wilson, I. a., Wyatt, R., Bewley, C. a., and Kwong, P. D. *Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. Science (New York, N.Y.)*, 317(5846):1930–4, 2007. ISSN 1095-9203. doi:10.1126/science.1145373. 38, 41

[50] Huang, C.-c., Tang, M., Zhang, M.-Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I. a., Wyatt, R., and Kwong, P. D. *Structure of a V3-containing HIV-1 gp120 core. Science (New York, N.Y.)*, 310(5750):1025–8, 2005. ISSN 1095-9203. doi:10.1126/science.1118398. 38

[51] Jackson, R. L., Busch, S. J., and Cardin, a. D. *Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes. Physiological reviews*, 71(2):481–539, 1991. ISSN 0031-9333. 49

[52] Jacobs, D. M., Saxena, K., Vogtherr, M., Bernado, P., Pons, M., and Fiebig, K. M. *Peptide binding induces large scale changes in inter-domain mobility in human Pin1. The Journal of biological chemistry*, 278(28):26174–82, 2003. ISSN 0021-9258. doi:10.1074/jbc.M300796200. 4

[53] Jakuschev, S., Ohlig, S., Farshi, P., Pickhinke, U., Boom, J. V. D., Ho, S., Dierker, T., Bordych, C., Grobe, K., Hoffmann, D., Dreier, R., Scho, H. R., van den Boom, J., Höing, S., and Schöler, H. R. *Sonic hedgehog shedding results in functional activation of the solubilized protein. Developmental cell*, 20(6):764–74, 2011. ISSN 1878-1551. doi:10.1016/j.devcel.2011.05.010. 76

[54] Johnson, Z., Proudfoot, a. E., and Handel, T. M. *Interaction of chemokines and glycosaminoglycans: a new twist in the regulation of chemokine function with opportunities for therapeutic intervention. Cytokine & growth factor reviews*, 16(6):625–36, 2005. ISSN 1359-6101. doi:10.1016/j.cytogfr.2005.04.006. 48, 66

[55] Jones, C. J., Beni, S., Limtiaco, J. F. K., Langeslay, D. J., and Larive, C. K. *Heparin characterization: challenges and solutions. Annual review of analytical chemistry (Palo Alto, Calif.)*, 4:439–65, 2011. ISSN 1936-1335. doi:10.1146/annurev-anchem-061010-113911. 49

[56] Jones, D. T., Buchan, D. W. a., Cozzetto, D., and Pontil, M. *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics (Oxford, England)*, 28(2):184–90, 2012. ISSN 1367-4811. doi:10.1093/bioinformatics/btr638. 18

[57] Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. *Template-based protein structure modeling using the RaptorX web server. Nature protocols*, 7(8):1511–22, 2012. ISSN 1750-2799. doi:10.1038/nprot.2012.085. 1

[58] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, a. a., Aflalo, C., and Vakser, I. a. *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proceedings of the National Academy of Sciences of the United States of America*, 89(6):2195–9, 1992. ISSN 0027-8424. 50, 57

[59] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature*, 181(4610):662–6, 1958. ISSN 0028-0836. 1

[60] KeÌĄkicheff, P., MarcÌĘelja, S., Senden, T. J., and Shubin, V. E. *Charge reversal seen in electrical double layer interaction of surfaces immersed in 2:1 calcium electrolyte. The Journal of Chemical Physics*, 99(8):6098, 1993. ISSN 00219606. doi:10.1063/1.465906. 53

[61] Khan, S., Gor, J., Mulloy, B., Perkins, S. J., and Mimms, S. *Semi-rigid solution structures of heparin by constrained X-ray scattering modelling: new insight into heparin-protein complexes. Journal of molecular biology*, 395(3):504–21, 2010. ISSN 1089-8638. doi:10.1016/j.jmb.2009.10.064. 49

[62] Klasse, P. J. *The molecular basis of HIV entry. Cellular microbiology*, 14(8):1183–92, 2012. ISSN 1462-5822. doi:10.1111/j.1462-5822.2012.01812.x. 38

[63] Koopmann, W. and Krangel, M. S. *Identification of a glycosaminoglycan-binding site in chemokine macrophage inflammatory protein-1alpha. The Journal of biological chemistry*, 272(15):10103–9, 1997. ISSN 0021-9258. 69, 71

[64] Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. *Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proceedings of the National Academy of Sciences of the United States of America*, 90(15):7176–80, 1993. ISSN 0027-8424. 18, 38

[65] Koziel, L., Kunath, M., Kelly, O. G., and Vortkamp, A. *Ext1-dependent heparan sulfate regulates the range of Ihh signaling during endochondral ossification. Developmental cell*, 6(6):801–13, 2004. ISSN 1534-5807. doi:10.1016/j.devcel.2004.05.009. 74

[66] Kuschert, G. S., Coulin, F., Power, C. a., Proudfoot, a. E., Hubbard, R. E., Hoogewerf, a. J., and Wells, T. N. *Glycosaminoglycans interact selectively with chemokines and modulate receptor binding and cellular responses. Biochemistry*, 38(39):12959–68, 1999. ISSN 0006-2960. 66

[67] Kwon, Y. D., Finzi, A., Wu, X., Dogo-Isonagie, C., Lee, L. K., Moore, L. R., Schmidt, S. D., Stuckey, J., Yang, Y., Zhou, T., Zhu, J., Vicic, D. a., Debnath, A. K., Shapiro, L., Bewley, C. a., Mascola, J. R., Sodroski, J. G., and Kwong, P. D. *Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. Proceedings of the National Academy of Sciences of the United States of America*, 109(15):5663–8, 2012. ISSN 1091-6490. doi:10.1073/pnas.1112391109. 38, 43

[68] Kwong, P. D., Wyatt, R., Majeed, S., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. a. *Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure (London, England : 1993)*, 8(12):1329–39, 2000. ISSN 0969-2126. 38

[69] Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. a. *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature*, 393(6686):648–59, 1998. ISSN 0028-0836. doi:10.1038/31405. 38

[70] Lapedes, A., Giraud, B., Liu, L., and Stormo, G. *Correlated Mutations in Models of Protein Sequences : Phylogenetic and Structural Effects. Lecture Notes-Monograph ...*, 33(1999):236–256, 1999. 18

[71] Lau, E. K., Paavola, C. D., Johnson, Z., Gaudry, J.-P., Geretti, E., Borlat, F., Kungl, A. J., Proudfoot, A. E., and Handel, T. M. *Identification of the glycosaminoglycan binding site of the CC chemokine, MCP-1: implications for structure and function in vivo. The Journal of biological chemistry*, 279(21):22294–305, 2004. ISSN 0021-9258. doi:10.1074/jbc.M311224200. 66

[72] Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology*, 487:545–74, 2011. ISSN 1557-7988. doi:10.1016/B978-0-12-381270-4.00019-6. 1

[73] Lee, S.-c., Guan, H.-h., Wang, C.-h., Huang, W.-n., Tjong, S.-c., Chen, C.-j., and Wu, W.-g. *Structural basis of citrate-dependent and heparan sulfate-mediated cell surface retention of cobra cardiotoxin A3. The Journal of biological chemistry*, 280(10):9567–77, 2005. ISSN 0021-9258. doi:10.1074/jbc.M412398200. 63

[74] Lesk, A. *Introduction to protein architecture: the structural biology of proteins.* Oxford University Press, Oxford, 1st edition, 2001. 28

[75] Li, D. and Roberts, R. *WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. Cellular and molecular life sciences : CMLS*, 58(14):2085–97, 2001. ISSN 1420-682X. 29

[76] Lietha, D., Chirgadze, D. Y., Mulloy, B., Blundell, T. L., and Gherardi, E. *Crystal structures of NK1-heparin complexes reveal the basis for NK1 activity and enable engineering of potent agonists of the MET receptor. The EMBO journal*, 20(20):5543–55, 2001. ISSN 0261-4189. doi:10.1093/emboj/20.20.5543. 63

[77] Liou, Y.-C., Zhou, X. Z., and Lu, K. P. *Prolyl isomerase Pin1 as a molecular switch to determine the fate of phosphoproteins. Trends in biochemical sciences*, 36(10):501–14, 2011. ISSN 0968-0004. doi:10.1016/j.tibs.2011.07.001. 3, 4

[78] Liu, D., Shriver, Z., Qi, Y., Venkataraman, G., and Sasisekharan, R. *Dynamic regulation of tumor growth and metastasis by heparan sulfate glycosaminoglycans. Seminars in*

thrombosis and hemostasis, 28(1):67–78, 2002. ISSN 0094-6176. doi:10.1055/s-2002-20565. 48

[79] Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G., and Subramaniam, S. *Molecular architecture of native HIV-1 gp120 trimers. Nature*, 455(7209):109–13, 2008. ISSN 1476-4687. doi:10.1038/nature07159. 37, 38, 41

[80] Liu, L., Cimbro, R., Lusso, P., and Berger, E. a. *Intraprotomer masking of third variable loop (V3) epitopes by the first and second variable loops (V1V2) within the native HIV-1 envelope glycoprotein trimer. Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20148–53, 2011. ISSN 1091-6490. doi:10.1073/pnas.1104840108. 37, 38, 42, 44

[81] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. *Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology*, 2012:251364, 2012. ISSN 1110-7251. doi:10.1155/2012/251364. 19, 47

[82] Lockless, S. W. *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science*, 286(5438):295–299, 1999. ISSN 00368075. doi:10.1126/science.286.5438.295. 18

[83] Lortat-Jacob, H., Grosdidier, A., and Imberty, A. *Structural diversity of heparan sulfate binding domains in chemokines. Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1229–34, 2002. ISSN 0027-8424. doi:10.1073/pnas.032497699. 68, 80

[84] Lu, K. P., Hanes, S. D., and Hunter, T. *A human peptidyl-prolyl isomerase essential for regulation of mitosis. Nature*, 380(6574):544–7, 1996. ISSN 0028-0836. doi:10.1038/380544a0. 3

[85] Madl, T., Güttler, T., Görlich, D., and Sattler, M. *Structural analysis of large protein complexes using solvent paramagnetic relaxation enhancements. Angewandte Chemie (International ed. in English)*, 50(17):3993–7, 2011. ISSN 1521-3773. doi:10.1002/anie.201007168. 6

[86] Mao, Y., Wang, L., Gu, C., Herschhorn, A., Désormeaux, A., Finzi, A., Xiang, S.-H., and Sodroski, J. G. *Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12438–43, 2013. ISSN 1091-6490. doi:10.1073/pnas.1307382110. 38

[87] Mao, Y., Wang, L., Gu, C., Herschhorn, A., Xiang, S.-H., Haim, H., Yang, X., and Sodroski, J. *Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. Nature structural & molecular biology*, 19(9):893–9, 2012. ISSN 1545-9985. doi:10.1038/nsmb.2351. 38

[88] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. a., Pagnani, A., Zecchina, R., and Sander, C. *Protein 3D structure computed from evolutionary sequence variation. PloS one*, 6(12):e28766, 2011. ISSN 1932-6203. doi:10.1371/journal.pone.0028766. 26

[89] Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. *Using information theory to search for co-evolving residues in proteins. Bioinformatics (Oxford, England)*, 21(22):4116–24, 2005. ISSN 1367-4803. doi:10.1093/bioinformatics/bti671. 18

[90] Matena, A., Sinnen, C., van den Boom, J., Wilms, C., Dybowski, J. N., Maltaner, R., Mueller, J. W., Link, N. M., Hoffmann, D., and Bayer, P. *Transient Domain Interactions Enhance the Affinity of the Mitotic Regulator Pin1 toward Phosphorylated Peptide Ligands. Structure (London, England : 1993)*, pages 1–9, 2013. ISSN 1878-4186. doi:10.1016/j.str.2013.07.016. 14

[91] MathWorks. *MATLAB R2013a*. The MathWorks Inc., Natick, Massachusetts, 2013. 25

[92] McLellan, J. S., Pancera, M., Carrico, C., Gorman, J., Julien, J.-P., Khayat, R., Louder, R., Pejchal, R., Sastry, M., Dai, K., O'Dell, S., Patel, N., Shahzad-ul Hussan, S., Yang, Y., Zhang, B., Zhou, T., Zhu, J., Boyington, J. C., Chuang, G.-Y., Diwanji, D., Georgiev, I., Kwon, Y. D., Lee, D., Louder, M. K., Moquin, S., Schmidt, S. D., Yang, Z.-Y., Bonsignori, M., Crump, J. a., Kapiga, S. H., Sam, N. E., Haynes, B. F., Burton, D. R., Koff, W. C., Walker, L. M., Phogat, S., Wyatt, R., Orwenyo, J., Wang, L.-X., Arthos, J., Bewley, C. a., Mascola, J. R., Nabel, G. J., Schief, W. R., Ward, A. B., Wilson, I. a., and Kwong, P. D. *Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. Nature*, 480(7377):336–43, 2011. ISSN 1476-4687. doi:10.1038/nature10696. 44

[93] McLellan, J. S., Yao, S., Zheng, X., Geisbrecht, B. V., Ghirlando, R., Beachy, P. A., and Leahy, D. J. *Structure of a heparin-dependent complex of Hedgehog and Ihog. Proceedings of the National Academy of Sciences of the United States of America*, 103(46):17208–13, 2006. ISSN 0027-8424. doi:10.1073/pnas.0606738103. 75, 85

[94] McLellan, J. S., Zheng, X., Hauk, G., Ghirlando, R., Beachy, P. a., and Leahy, D. J. *The mode of Hedgehog binding to Ihog homologues is not conserved across different phyla. Nature*, 455(7215):979–83, 2008. ISSN 1476-4687. doi:10.1038/nature07358. 75

[95] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. *Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–301, 2011. ISSN 1091-6490. doi:10.1073/pnas.1111471108. 2, 17, 18, 19, 22, 24, 25, 26, 28, 29, 32, 34

[96] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of computational chemistry*, 30(16):2785–91, 2009. ISSN 1096-987X. doi:10.1002/jcc.21256. 1

[97] Morrison, K. L. and Weiss, G. A. *Combinatorial alanine-scanning. Current opinion in chemical biology*, 5(3):302–7, 2001. ISSN 1367-5931. 61

[98] Neher, E. *How frequent are correlated changes in families of protein sequences? Proceedings of the National Academy of Sciences of the United States of America*, 91(1):98–102, 1994. ISSN 0027-8424. 18

[99] Ohlig, S., Pickhinke, U., Sirko, S., Bandari, S., Hoffmann, D., Dreier, R., Farshi, P., Goetz, M., and Grobe, K. *An emerging role of Sonic hedgehog shedding as a modulator of heparan sulfate interactions. The Journal of biological chemistry*, 287(52):1–24, 2012. ISSN 1083-351X. doi:10.1074/jbc.M112.356667. 74, 82

[100] Oliphant, T. E. *Python for Scientific Computing. Computing in Science & Engineering*, 9(3):10–20, 2007. ISSN 1521-9615. doi:10.1109/MCSE.2007.58. 25, 56

[101] Palm, W., Swierczynska, M. M., Kumari, V., Ehrhart-bornstein, M., Bornstein, S. R., and Eaton, S. *Secretion and Signaling Activities of Lipoprotein-Associated Hedgehog and Non-Sterol-Modified Hedgehog in Flies and Mammals. PLoS Biology*, 11(3):e1001505, 2013. ISSN 1545-7885. doi:10.1371/journal.pbio.1001505. 74

[102] Pancera, M., Majeed, S., Ban, Y.-E. A., Chen, L., Huang, C.-c., Kong, L., Kwon, Y. D., Stuckey, J., Zhou, T., Robinson, J. E., Schief, W. R., Sodroski, J., Wyatt, R., and Kwong, P. D. *Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. Proceedings of the National Academy of Sciences of the United States of America*, 107(3):1166–71, 2010. ISSN 1091-6490. doi:10.1073/pnas.0911004107. 38, 41, 45

[103] Pancera, M., Shahzad-Ul-Hussan, S., Doria-Rose, N. a., McLellan, J. S., Bailer, R. T., Dai, K., Loesgen, S., Louder, M. K., Staupe, R. P., Yang, Y., Zhang, B., Parks, R.,

Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Amin, M. N., Wang, L.-X., Burton, D. R., Koff, W. C., Nabel, G. J., Mascola, J. R., Bewley, C. a., and Kwong, P. D. *Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. Nature structural & molecular biology*, 20(7):804–13, 2013. ISSN 1545-9985. doi:10.1038/nsmb.2600. 44

[104] Pastorino, L., Sun, A., Lu, P.-J., Zhou, X. Z., Balastik, M., Finn, G., Wulf, G., Lim, J., Li, S.-H., Li, X., Xia, W., Nicholson, L. K., and Lu, K. P. *The prolyl isomerase Pin1 regulates amyloid precursor protein processing and amyloid-beta production. Nature*, 440(7083):528–34, 2006. ISSN 1476-4687. doi:10.1038/nature04543. 4

[105] Pepinsky, R. B., Rayhorn, P., Day, E. S., Dergay, a., Williams, K. P., Galdes, a., Taylor, F. R., Boriack-Sjodin, P. a., and Garber, E. a. *Mapping sonic hedgehog-receptor interactions by steric interference. The Journal of biological chemistry*, 275(15):10995–1001, 2000. ISSN 0021-9258. 75, 76, 82

[106] Pepinsky, R. B., Zeng, C., Wen, D., Rayhorn, P., Baker, D. P., Williams, K. P., Bixler, S. a., Ambrose, C. M., Garber, E. a., Miatkowski, K., Taylor, F. R., Wang, E. a., and Galdes, a. *Identification of a palmitic acid-modified form of human Sonic hedgehog. The Journal of biological chemistry*, 273(22):14037–45, 1998. ISSN 0021-9258. 74

[107] Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. T. *Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-âĎń. Resolution, Obtained by X-Ray Analysis. Nature*, 185(4711):416–422, 1960. ISSN 0028-0836. doi:10.1038/185416a0. 1

[108] Poon, A. and Chao, L. *The rate of compensatory mutation in the DNA bacteriophage phiX174. Genetics*, 170(3):989–99, 2005. ISSN 0016-6731. doi:10.1534/genetics.104.039438. 18

[109] Porter, J. A., Young, K. E., and Beachy, P. A. *Cholesterol Modification of Hedgehog Signaling Proteins in Animal Development. Science*, 274(5285):255–259, 1996. ISSN 0036-8075. doi:10.1126/science.274.5285.255. 74

[110] Procaccini, A., Lunt, B., Szurmant, H., Hwa, T., and Weigt, M. *Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PloS one*, 6(5):e19729, 2011. ISSN 1932-6203. doi:10.1371/journal.pone.0019729. 19, 22, 24, 26

[111] Rabenstein, D. L. *Heparin and heparan sulfate: structure and function. Natural Product Reports*, 19(3):312–331, 2002. ISSN 02650568. doi:10.1039/b100916h. 49

[112] Rajarathnam, K., Sykes, B. D., Kay, C. M., Dewald, B., Geiser, T., Baggiolini, M., and Clark-Lewis, I. *Neutrophil activation by monomeric interleukin-8. Science (New York, N.Y.)*, 264(5155):90–2, 1994. ISSN 0036-8075. 66

[113] Ranganathan, R., Lu, K. P., Hunter, T., and Noel, J. P. *Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. Cell*, 89(6):875–86, 1997. ISSN 0092-8674. 3, 4, 5, 13

[114] Ren, M., Guo, Q., Guo, L., Lenz, M., Qian, F., Koenen, R. R., Xu, H., Schilling, A. B., Weber, C., Ye, R. D., Dinner, A. R., and Tang, W.-J. *Polymerization of MIP-1 chemokine (CCL3 and CCL4) and clearance of MIP-1 by insulin-degrading enzyme. The EMBO journal*, 29(23):3952–66, 2010. ISSN 1460-2075. doi:10.1038/emboj.2010.256. 66, 67, 71

[115] Roy, A., Kucukural, A., and Zhang, Y. *I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols*, 5(4):725–38, 2010. ISSN 1750-2799. doi:10.1038/nprot.2010.5. 1

[116] Rubin, J. B., Choi, Y., and Segal, R. A. *Cerebellar proteoglycans regulate sonic hedgehog responses during development. Development (Cambridge, England)*, 129(9):2223–32, 2002. ISSN 0950-1991. 74

[117] Rusert, P., Krarup, A., Magnus, C., Brandenberg, O. F., Weber, J., Ehlert, A.-K., Regoes, R. R., Günthard, H. F., and Trkola, A. *Interaction of the gp120 V1V2 loop with a neighboring gp120 unit shields the HIV envelope trimer against cross-neutralizing antibodies. The Journal of experimental medicine*, 208(7):1419–33, 2011. ISSN 1540-9538. doi:10.1084/jem.20110196. 37, 38, 42, 44

[118] Ryan, K. E. and Chiang, C. *Hedgehog secretion and signal transduction in vertebrates. The Journal of biological chemistry*, 287(22):17905–13, 2012. ISSN 1083-351X. doi:10.1074/jbc.R112.356006. 74

[119] Sali, A. and Blundell, T. L. *Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology*, 234(3):779–815, 1993. ISSN 0022-2836. doi:10.1006/jmbi.1993.1626. 41, 63, 76

[120] Schlessinger, J., Plotnikov, A. N., Ibrahimi, O. A., Eliseenkova, A. V., Yeh, B. K., Yayon, A., Linhardt, R. J., and Mohammadi, M. *Crystal structure of a ternary*

FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization. Molecular cell, 6(3):743–50, 2000. ISSN 1097-2765. 63

[121] Schmid, F. X. and Baldwin, R. L. *Acid catalysis of the formation of the slow-folding species of RNase A: evidence that the reaction is proline isomerization. Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4764–8, 1978. ISSN 0027-8424. 4

[122] Schmid, F. X. and Baldwin, R. L. *The rate of interconversion between the two unfolded forms of ribonuclease A does not depend on guanidinium chloride concentration. Journal of molecular biology*, 133(2):285–7, 1979. ISSN 0022-2836. 4

[123] Schrödinger, L. *The PyMOL Molecular Graphics System Version 1.3.x*, 2010. 43

[124] Schüttelkopf, A. W. and van Aalten, D. M. F. *PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta crystallographica. Section D, Biological crystallography*, 60(Pt 8):1355–63, 2004. ISSN 0907-4449. doi:10.1107/S0907444904011679. 57, 58

[125] Sharma, A., Askari, J. A., Humphries, M. J., Jones, E. Y., and Stuart, D. I. *Crystal structure of a heparin- and integrin-binding segment of human fibronectin. The EMBO journal*, 18(6):1468–79, 1999. ISSN 0261-4189. doi:10.1093/emboj/18.6.1468. 63

[126] Shaya, D., Zhao, W., Garron, M.-L., Xiao, Z., Cui, Q., Zhang, Z., Sulea, T., Linhardt, R. J., Cygler, M., Engineering, B., and Studies, I. *Catalytic mechanism of heparinase II investigated by site-directed mutagenesis and the crystal structure with its substrate. The Journal of biological chemistry*, 285(26):20051–61, 2010. ISSN 1083-351X. doi:10.1074/jbc.M110.101071. 63

[127] Smith, T. F., Gaitatzes, C., Saxena, K., and Neer, E. J. *The WD repeat: a common architecture for diverse functions. Trends in biochemical sciences*, 24(5):181–5, 1999. ISSN 0968-0004. 29

[128] Starcich, B. R., Hahn, B. H., Shaw, G. M., McNeely, P. D., Modrow, S., Wolf, H., Parks, E. S., Parks, W. P., Josephs, S. F., and Gallo, R. C. *Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell*, 45(5):637–48, 1986. ISSN 0092-8674. 40

[129] Stringer, S. E., Forster, M. J., Mulloy, B., Bishop, C. R., Graham, G. J., and Gallagher, J. T. *Characterization of the binding site on heparan sulfate for macrophage inflammatory protein 1alpha. Blood*, 100(5):1543–50, 2002. ISSN 0006-4971. 66

[130] Taylor, W. R. and Hatrick, K. *Compensating changes in protein multiple sequence alignments. Protein engineering*, 7(3):341–8, 1994. ISSN 0269-2139. 18

[131] Team, R. D. C. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. 61, 78

[132] The, I., Bellaiche, Y., and Perrimon, N. *Hedgehog movement is regulated through tout velu-dependent synthesis of a heparan sulfate proteoglycan. Molecular cell*, 4(4):633–9, 1999. ISSN 1097-2765. 74

[133] Tran, E. E. H., Borgnia, M. J., Kuybeda, O., Schauder, D. M., Bartesaghi, A., Frank, G. a., Sapiro, G., Milne, J. L. S., and Subramaniam, S. *Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. PLoS pathogens*, 8(7):e1002797, 2012. ISSN 1553-7374. doi:10.1371/journal.ppat.1002797. 37, 38, 41

[134] Verdecia, M. A., Bowman, M. E., Lu, K. P., Hunter, T., and Noel, J. P. *Structural basis for phosphoserine-proline recognition by group IV WW domains. Nature structural biology*, 7(8):639–43, 2000. ISSN 1072-8368. doi:10.1038/77929. 4

[135] Vyas, N., Goswami, D., Manonmani, A., Sharma, P., Ranganath, H. a., Vijayraghavan, K., Shashidhara, L. S., Sowdhamini, R., and Mayor, S. *Nanoscale organization of hedgehog is essential for long-range signaling. Cell*, 133(7):1214–27, 2008. ISSN 1097-4172. doi:10.1016/j.cell.2008.05.026. 74

[136] Wang, Y., Rawi, R., Wilms, C., Heider, D., Yang, R., and Hoffmann, D. *A small set of succinct signature patterns distinguishes Chinese and non-Chinese HIV-1 genomes. PloS one*, 8(3):e58804, 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0058804. 45

[137] Wedemeyer, W. J., Welker, E., and Scheraga, H. A. *Proline cis-trans isomerization and protein folding. Biochemistry*, 41(50):14637–44, 2002. ISSN 0006-2960. 3

[138] Wei, B., Han, N., Liu, H.-z., Rayner, A., and Rayner, S. *Use of mutual information arrays to predict coevolving sites in the full length HIV gp120 protein for subtypes B and C. Virologica Sinica*, 26(2):95–104, 2011. ISSN 1995-820X. doi:10.1007/s12250-011-3188-7. 38

[139] Weigt, M., White, R. a., Szurmant, H., Hoch, J. a., and Hwa, T. *Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences of the United States of America*, 106(1):67–72, 2009. ISSN 1091-6490. doi:10.1073/pnas.0805923106. 17

[140] Weissenhorn, W., Dessen, A., Harrison, S. C., Skehel, J. J., and Wiley, D. C. *Atomic structure of the ectodomain from HIV-1 gp41.* *Nature*, 387(6631):426–30, 1997. ISSN 0028-0836. doi:10.1038/387426a0. 38

[141] Wells, T. N. C., Power, C. a., Shaw, J. P., and Proudfoot, A. E. I. *Chemokine blockers–therapeutics in the making?* *Trends in pharmacological sciences*, 27(1):41–7, 2006. ISSN 0165-6147. doi:10.1016/j.tips.2005.11.001. 66

[142] Whalen, D. M., Malinauskas, T., Gilbert, R. J. C., and Siebold, C. *Structural insights into proteoglycan-shaped Hedgehog signaling.* *Proceedings of the National Academy of Sciences*, 2013:1–6, 2013. ISSN 0027-8424. doi:10.1073/pnas.1310097110. 82, 87

[143] White, T. a., Bartesaghi, A., Borgnia, M. J., Meyerson, J. R., de la Cruz, M. J. V., Bess, J. W., Nandwani, R., Hoxie, J. a., Lifson, J. D., Milne, J. L. S., and Subramaniam, S. *Molecular architectures of trimeric SIV and HIV-1 envelope glycoproteins on intact viruses: strain-dependent variation in quaternary structure.* *PLoS pathogens*, 6(12):e1001249, 2010. ISSN 1553-7374. doi:10.1371/journal.ppat.1001249. 38

[144] Wilson, D. K., Cerna, D., and Chew, E. *The 1.1-angstrom structure of the spindle checkpoint protein Bub3p reveals functional regions.* *The Journal of biological chemistry*, 280(14):13944–51, 2005. ISSN 0021-9258. doi:10.1074/jbc.M412919200. 29

[145] Witt, D. P. and Lander, a. D. *Differential binding of chemokines to glycosaminoglycan subpopulations.* *Current biology : CB*, 4(5):394–400, 1994. ISSN 0960-9822. 48, 66

[146] Wolpert, L. *Positional information and the spatial pattern of cellular differentiation.* *Journal of theoretical biology*, 25(1):1–47, 1969. ISSN 0022-5193. 74

[147] Wu, S.-R., Löving, R., Lindqvist, B., Hebert, H., Koeck, P. J. B., Sjöberg, M., and Garoff, H. *Single-particle cryoelectron microscopy analysis reveals the HIV-1 spike as a tripod structure.* *Proceedings of the National Academy of Sciences of the United States of America*, 107(44):18844–9, 2010. ISSN 1091-6490. doi:10.1073/pnas.1007227107. 38

[148] Yaffe, M. B. *Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism.* *Science*, 278(5345):1957–1960, 1997. ISSN 00368075. doi:10.1126/science.278.5345.1957. 3

[149] Yanofsky, C., Horn, V., and Thorpe, D. *Protein Structure Relationships Revealed By Mutational Analysis.* *Science (New York, N.Y.)*, 146(3651):1593–4, 1964. ISSN 0036-8075. 18

[150] Yeh, E. S. and Means, A. R. *PIN1, the cell cycle and cancer. Nature reviews. Cancer*, 7(5):381–8, 2007. ISSN 1474-175X. doi:10.1038/nrc2107. 4

[151] Zhang, F., McLellan, J. S., Ayala, A. M., Leahy, D. J., and Linhardt, R. J. *Kinetic and structural studies on interactions between heparin or heparan sulfate and proteins of the hedgehog signaling pathway. Biochemistry*, 46(13):3933–41, 2007. ISSN 0006-2960. doi:10.1021/bi6025424. 78, 82, 87

[152] Zhang, Y. *Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins*, 2013. ISSN 1097-0134. doi:10.1002/prot.24341. 1

[153] Zhang, Y., Daum, S., Wildemann, D., Zhou, X. Z., Verdecia, M. a., Bowman, M. E., Lücke, C., Hunter, T., Lu, K.-P., Fischer, G., and Noel, J. P. *Structural basis for high-affinity peptide inhibition of human Pin1. ACS chemical biology*, 2(5):320–8, 2007. ISSN 1554-8937. doi:10.1021/cb7000044. 4

[154] Zhu, P., Winkler, H., Chertova, E., Taylor, K. a., and Roux, K. H. *Cryoelectron tomography of HIV-1 envelope spikes: further evidence for tripod-like legs. PLoS pathogens*, 4(11):e1000203, 2008. ISSN 1553-7374. doi:10.1371/journal.ppat.1000203. 37

List of Publications

Publications

Peer-reviewed

Yan Wang, Reda Rawi, **Christoph Wilms**, Dominik Heider, Rongge Yang and Daniel Hoffmann. *A Small Set of Succinct Signature Patterns Distinguishes Chinese and Non-Chinese HIV-1 Genomes.* PLOS One, 2013

Anja Matena, Christian Sinnen, Johannes van den Boom, **Christoph Wilms**, J. Nikolaj Dybowski, Ricarda Maltaner, Jonathan W. Mueller, Nina M. Link, Daniel Hoffmann, and Peter Bayer. *Transient Domain Interactions Enhance the Affinity of the Mitotic Regulator Pin1 toward Phosphorylated Peptide Ligands.* Structure, 2013

Submitted

Dominik Heider, Nikolaj Dybowski, **Christoph Wilms** and Daniel Hoffman. *A simple structure-based model for the prediction of HIV-1 co-receptor tropism.* BMC Bioinformatics, 2013

Rocio Rebollido-Rios, **Christoph Wilms**, Stanislav Jakuschev, Andrea Vortkamp, Kay Grobe and Daniel Hoffmann. *Signaling Domain of Sonic Hedgehog as Cannibalistic Calcium-Regulated Zinc-Peptidase.* PLOS Computational Biology, 2013

Acknowledgements

After three really enjoyable and interesting years this chapter of my life has come to an end. At this point I would like to thank everyone who contributed to this great experience.

First of all, I would like to thank Daniel Hoffmann for providing me the opportunity to start and complete this PhD thesis. It was a time full of challenging and interesting scientific problems, where I always knew I could rely on his shear endless wisdom.

Second of all, I would like to thank all my colleagues from the Bioinformatics Department of the University Duisburg-Essen, especially Jan Taubenheim for his help with all my biological-related problems and Reda Rawi for all the inspiring discussions.

Additionally, I would like to thank Vera Kleptkova and Martin Peters for reading this work and providing me with useful feedback.

Furthermore, I would like to thank my friends and family. Without the support of my family and especially my mother throughout my life I would not have managed to get this far.

Last but not least, I would like to thank my wife Jasmin for her continuous loving support.

Declarations

Erklärung:

Hiermit erkläre ich, gem. § 6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik zur Erlangung der Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema „Methods for the Prediction of Complex Biomolecular Structures“ zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Christoph Wilms befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen den, _____ _____

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) c) + e) der Promotionsordnung Fakultäten für Biologie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

Stein a. d. Traun den, _____ _____

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Stein a. d. Traun den, _____ _____